Philosophische Analyse/ Philosophical Analysis

Herausgegeben von/Edited by
Rafael Hüntelmann, Christian Kanzian, Uwe Meixner,
Richard Schantz, Erwin Tegtmeier

Band/Volume 73

From Arithmetic to Metaphysics

A Path through Philosophical Logic

Edited by
Ciro de Florio and Alessandro Giordani

DE GRUYTER

ISBN 978-3-11-052882-4
e-ISBN (PDF) 978-3-11-052949-4
e-ISBN (EPUB) 978-3-11-052901-2
ISSN 2198-2066

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

Printing: CPI books GmbH, Leck
♾ Printed on acid-free paper
Printed in Germany

www.degruyter.com

to Sergio Galvan

Contents

Preface

The distinction between *appearance* and *reality*, introduced by Parmenides and Heraclitus, is said to be one of the pillars of Western thought. And this is precisely what this volume *is*, i.e. a real tribute to its addressee, even though it does not *appear* as the typical collection of papers in honor of someone. It is indeed an homage to a professor and to his influence in the research areas which the collected essays are a representation of.

Sergio Galvan graduated in 1969 under Evandro Agazzi's tuition with a dissertation on Alfred Tarski's semantic conception of truth. At that time Alfred Tarski was still alive and his well-known essay *Truth and Proof* appeared on Scientific American. Galvan spent a period as visiting researcher in Germany; then, he taught at secondary and high school, bravely explaining to – likely astounded – students the theory of syllogism. Finally, he worked at the universities in Milan and Verona, becoming full professor of Logic at Trento University in 1994. Three years later, he returned to the Catholic University of Milan where he previously had studied. At Catholic University he was chair of Logic, Philosophy of Science and Analytic Ontology.

In the Sixties, the Italian philosophical scenario, in which Galvan intellectually grew up, was dominated by discussions concerning political philosophy (in particular, the debate between Catholics and Communists), philosophy of existence, aesthetics, and moral philosophy. Logic and philosophy have been just introduced, each of them in his own way and peculiar attitude, by such scholars as Ludovico Geymonat, Ettore Casari, Alberto Pasquinelli, and Galvan's teacher Evandro Agazzi,. Metaphysics lied idle, rather forgotten. However, Catholic University still paid great attention to the science of being, thanks to some important teachers: Amato Masnovo, Gustavo Bontadini and Sofia Vanni Rovighi. Indeed, Gustavo Bontadini was the former teacher of Evandro Agazzi and Galvan himself attended his lectures. Throughout the years of study, Sergio nurtured the passion for the classic metaphysical themes – from the problem of becoming to the ground of reality, from the constitution of entities to the problem of universals – and the interest for them has been a constant in his thought.

Sergio Galvan is a logician and all his scientific work embodies Hilbert's motto: clear thought is axiomatic thought. His main areas of research can be summarized as follows: mathematical logic, modal logic, logic of explanation and metaphysics. His first field of research concerns the analysis of Tarskian theory of truth and its connection with the classical correspondentist conception of truth. The passage from the study of Tarski's theory of truth to the general limitations of axiomatic systems was quick and leaded Galvan to face the problem of formalizing

DOI: 10.1515/9783110529494-001

arithmetics. The outcome of these years of work, *Formal theory of natural numbers* (in Italian: *Teoria formale dei numeri naturali*), is a formidable book on first and second order, and continues to be an unsurpassed landmark for the complete presentation of proposed results. *Introduction to the Incompleteness Theorems* (in Italian: *Introduzione ai teoremi di incompletezza*; in 2006 the book was translated in German as: *Einführung in die Unvollständigkeitstheoreme*) is the synthesis of many courses on Gödel's theorem and on its philosophical meaning which Galvan has been taught for three decades. In the last years Galvan investigated the analysis of weak systems of arithmetic, from Q to PA, correlating different versions of induction axioms with their epistemic commitment towards forms of more or less finitary construction, taking part in the debate on finitism that was raised, among others, by William Tait and Charles Parsons. According to Galvan, the reflection on axiomatic theories has both an intrinsic value – related to the dynamic of logical characterization – and a fundamental philosophical relevance, which is an aspect he has been always interested in: limitation theorems (Gödel, Tarski, Löwenheim Skolem, Church) can be widely applied to the human knowledge, the realism, the nature of eidetic intuition, and even in the conception of the mechanic thought.

Galvan's second vast area of research is constituted by the modal logic, a subject he helped to introduce to Italy, with his textbook *Intensional Logics. Systems of modal, denotic, and epistemic propositional logic* (in Italian: *Logiche intensionali, Sistemi proposizionali di logica modale, deontica, epistemica*). In particular, his essays on epistemic and deontic laws, with reference to the formalization of Hume's law and Kantian deontic principles, are worth of consideration. Moreover, Galvan has always tightly connected the investigations about modal logic with the metaphysical reflection; a specific interpretation of the semantics of modal logic brought him to a construal of the negation, analyzed in the book *Non contradiction and Excluded Middle* (in Italian: *Non contraddizione e terzo escluso*). Moreover, he formulated a modal system for the logic of essence, PIES5, that aims to adequately grasp the classical conception of modalities and the essential constitution of individuals in order to come up with an original solution to the trans-world identity and persistence problems.

Galvan's view of metaphysics, reflects, in a way his classical philosophical background (in a word, the Aristotelian and Thomistic thinking) but it is enriched by some fundamental issues of the philosophy of science: metaphysics has, at the end of the day, the ultimate explanatory function of the experience. Accordingly, the issue of explanation has been always relevant in Galvan's thought, particularly as far as the finalistic explanations occurring in the human sciences are concerned.

From Arithmetic to Metaphysics is in no way just the starting point or the final destination of a journey; this collection of essays rather represents the poles of

philosophical activity that has been lasting for over half a century. And it does not stop even today. As we speak, Sergio Galvan is writing a co-authored book (together with Paolo Mancosu and Richard Zach) on an exhaustive analysis of the *cut* theorem, the core of proof theory. Meanwhile, Galvan is developing his view on the metaphysics of possibility, within the system PIES5 inspired to Francisco Suarez' work.

Sergio Galvan's several intellectual virtues can be probably synthetized in a double definition that could be labeled as 'internally unstable': Sergio is gifted by an insatiable intellectual curiosity bound to a deep systematic nature that rejects any rash and superficial hypothesis. His idea is that reality is a plural, multi-faced and irreducibly complex whole. He has always been skeptical about the rhetorical chitchat of much contemporary philosophy, acting sometimes as a harsh critic of the exaggerations in analytical philosophy. He does not especially approve the excessive specialization in philosophical questions, caused by an ongoing widening and splitting of the research areas, leading to the loss of a systematic and unitary vision inherited by the classical metaphysics. Therefore, he is quite skeptical about the possibility of a synthesis coming after any thematic investigation, and does not believe that philosophical research should boil down to an activity of «puzzle-solving», no matter how acute and brain-teasing.

On the contrary, he firmly maintains that philosophy pursues a major objective: to provide a coherent picture of the world and of our place in it. That is the reason why he has always admired two scholars that, despite their differences, strove towards a unitary framework of knowledge: Franz von Kutschera, who honored us with an original contribution to the present volume and Jonathan Lowe, who prematurely passed away, but would have certainly had something precious to say in the following pages.

As we said at the beginning, the collected essays are independent and yet go through all the topics Sergio Galvan cares for: from the Philosophy of mathematics and Logic to the Philosophy of science up to Metaphysics and Philosophy of religion. The authors are colleagues and friends who has always admired Sergio's deepest intellectual honesty and his passion for being a philosopher. However, Sergio is a mountaineer (he was born in Levico Terme, at the foot of Dolomites) and therefore not very inclined to overwhelming compliments and praises. A glass of wine – a good one, though – hopefully will serve the same.

Ciro De Florio
Alessandro Giordani

Michele Abrusci

Hilbert's τ and ϵ in Proof Theory: a proof-theoretical representation of universal and existential statements

Abstract: In section 1, I expose in an informal way the rules – and the logical rules – on the proofs of the universal statements and existential statements, and the rules – and the logical rules – on the deductions from these statements. In section 2, I show how Hilbert's operators τ and ϵ allow a representation of the universal statements and existential statements which is strictly related to the logical rules on the proofs of these statements and to the logical rules on the deductions from these statements, so that we may say that Hilbert in the introduction of the operators τ and ϵ aimed to propose a kind of *proof-theoretical* representation of the universal statements and existential statements. In section 3, I show the logical naturalness and the logical depth of this representation of universal and existential statements, since τ-axiom and ϵ-axiom – which are the implicit definitions of these operators – arise in a very natural way from a deep analysis of what happens when we try to prove (in sequent calculus) the sequents $\forall xA \vdash \forall xA$ and $\exists xA \vdash \exists xA$ from the identity axiom $A \vdash A$.

1 Universal and existential statements: rules on proofs and deductions

In his talks Hilbert (1923) and Hilbert (1926) devoted to the foundations of mathematics, Hilbert introduced the operators τ and ϵ in order to obtain a representation of the universal statements and existential statements different from the Frege's representation of these statements.

Hilbert aimed to use this new representation firstly for the universal statements and existential statements belonging to mathematical analysis (because in these lectures he proposed the formalization of mathematical analysis by using the operators τ and ϵ), but also for the universal statements and existential statements belonging to the ordinary language (because in these lectures he proposed examples of universal and existential statements coming from the ordinary language, and showed how these statements may be well represented by means of these operators). Therefore, it is not true that Hilbert has been interested

Michele Abrusci: University of Rome 3

DOI: 10.1515/9783110529494-002

exclusively in mathematical statements (since the examples shown by Hilbert are also universal and existential statements not belonging to mathematics) and it is not true that Hilbert was interested primarily in the representation of first-order arithmetical statements (since mathematical analysis contains statements that are not first-order statements).

In the section 2 I will show that the representation of the universal statements and the existential statements proposed by Hilbert by means of the operators τ and ϵ is very close to the logical rules on proofs of universal and existential statements and the logical rules on deductions from universal and existential statements.

In this section 1, I will expose the basic notions used in section 2, and in particular the logical rules on the proofs of the universal statements and the existential statements and the logical rules on the deductions from the universal statements and the existential statements:

- in 1.1 I will fix what are the universal statements and the existential statements, and to specify what are the (universal, existential) mathematical statements;
- in 1.3 and 1.4 I will expose the *rules* (and in particular the *logical rules*) on the proofs of the universal statements and the existential statements, and the *rules* (and in particular the *logical rules*) on the deductions from the universal statements and the existential statements (the exposition of these rules is largely inspired by the investigations done in the proof-theory of last century);
- in 1.2 I will consider the notion of *generic objects inside proofs and deductions*, whose role is important in the logical rules on the proofs of the universal statements and the existential statements and on the deductions from the universal statements and the existential statements.

In this paper I will deal with the concepts *proofs* and *deductions* without the reference to a particular formalism (and so in a way to be applied to any formalism), by using the following principles:

- a *proof* of a statement A, i.e. a proof whose conclusion is A, allows to accept A i.e. (in classical logic) to discover its truth;
- a *deduction* from a statement A, i.e. a deduction whose premise is A, allows to infer something from the statement A i.e. (in classical logic) to discover something that would be true if A is true;
- *proofs* and *deductions* are always concrete and finite objects;
- given a proof π of a statement A, the other conclusions of π and the premisses of π are the *context of the proof* π of A; given a deduction π from a statement A, the other premises of π and the conclusions of π are the *context of the deduction* π from A;

- a *proof* of a statement A *in a context* Γ allows to accept A (i.e. in classical logic to discover the truth of A) when the context Γ holds (i.e. in classical logic when the premises are true and the other conclusions are false);
- a *deduction* from a statement A *in a context* Γ allows to infer something from A when the context Γ holds (i.e. to discover the truth of a conclusion when A and the other premises are true and the other conclusions are false);
- a *deduction* of a statement B from a statement A (in a context Γ) is *a proof* of $A \to B$ (in the same context Γ), and a *proof* of $A \to B$ (in the context Γ) is a deduction of B from A (in the same context Γ).

1.1 Universal statements and existential statements

Each universal or existential statement refers to a *type*, i.e. to a class or a set: indeed, any *universal statement* is an universal statement on some type T, and any *existential statement* is an existential statement on some type T. Examples of types: the type of natural numbers, the type of propositions, the type of functions on a given type, the type of predicates on a given type, etc.

Let T be an arbitrary type:

- *universal statements on the type* T are the statements that may be expressed in the form

$$A[x] \text{ for every } x \text{ of type } T$$

- *existential statements on the type* T are the statements that may be expressed in the form

$$A[x] \text{ for some } x \text{ of type } T.$$

A *mathematical* type is a type of mathematical objects (e.g. the type of natural numbers), or a type of functions on a type of mathematical objects, or a type of predicates on a type of mathematical objects.

An universal statement is *mathematical* when it is an universal statement on a mathematical type T, and an existential statement is *mathematical* when it is an existential statement on a mathematical type T

A *first-order statement* is a statement where all the universal and existential statements concern the same type T of objects; so, a first-order statement does not contain universal statements or existential statements on a type of propositions or a on a type of functions or an a type of predicates.

1.2 Generic objects inside proofs and deductions

Generic objects inside a proof, and generic objects inside a deduction, play an important role in the logical rules on the proofs of universal statements and existential statements and in the logical rules on the deductions from universal statements and existential statements.

Generic objects inside proofs and deductions may be defined as follows, in a way inspired by the developments of logic in the last century.

- Let π be a proof of a statement $A[a]$ where a is an object of a type T: if in π on the object a is used only what follows from the fact that a is an object of type T (i.e. no particular property of a is used in π), we say that a is *a generic object of type T inside the proof* π of $A[a]$.
- Let π be a deduction from a statement $A[a]$ where a is an object of a type T: if in π on the object a is used only what follows from the fact that a is an object of type T such that $A[a]$ holds (i.e. no particular property of a is used in π), we say that a is *a generic object of type T inside the deduction* π from $A[a]$.

In particular,

- if x is a variable of type T and π is a logical proof of $A[x]$ where x does not occur in the context, then x is a generic object of type T in the logical proof π of $A[x]$;
- if x is a variable of type T and π is a logical deduction from $A[x]$ where x does not occur in the context, then x is a generic object of type T in the logical deduction π.

When an object a of type T is a generic object of type T *inside a proof of a statement* $A[a]$ or *inside a deduction from a statement* $A[a]$, this does not mean that a is *a generic object at all* , i.e. an object that has only the properties common to all the objects of type T: simply we say that such an object a is *considered as a generic object* inside that particular proof of $A[a]$ or inside that particular deduction from $A[a]$.

When inside a proof (or a deduction) π an object a of type T is a generic object of type T, π is a *schema* of a class of proofs (a class of deductions), containing a proof (a deduction) for each object of T, obtained in an *uniform* way from π. Indeed:

- if π is a proof of a statement $A[a]$ where a is a generic object of type T, then we may produce a proof of $A[t]$ for each object b of T in an *uniform way* i.e. in a way not depending on particular features of b, as follows: simply, replace b for a everywhere in the proof of $A[a]$, i.e. consider b as a generic object of type T in π;

– if π is a deduction from a statement $A[a]$ where a is a generic object of type T, we may produce a deduction from $A[t]$ for each object b of T in an *uniform way* i.e. in a way not depending on particular features of b, as follows: simply, replace b for a everywhere in the deduction from $A[a]$, i.e. consider b as a generic object of type T in π.

So, when π is a proof (or a deduction) where a is a generic object of type T, π may be called a *uniform proof (uniform deduction)* or a *parametric proof (parametric deduction)*.

Uniform proofs are very important in mathematical research. An uniform proof has been the tool used by Girard in Girard (1971) to obtain the strong normalization theorem for the *System F* by means of the *Reducibility candidates* (a kind of generic object), and so avoiding the impredicativities.

Uniform proofs (parametric proofs) are very important *in computer science*: to the uniform proofs (according the Curry-Howard correspondence Howard (1980)) correspond *uniform programs*, i.e. programs which give in an uniform way the value for each input (and so independently from the given input).

1.3 Universal and existential statements: logical rules on proofs and deductions

A way to prove universal or existential statements and to deduce from universal or existential statements may be denominated a *logical rule* when it does not depend on a given particular type and may be applied to any type.

We may say that – form a logical point of view – the universal statements and the existential statements are defined by the logical rules on the proofs of these statements and on the deductions from these statements, and by the general logical rules concerning the statements. These rules are well formalized in the sequent calculus introduced by Gentzen in Gentzen (1935).

The following are the *logical rules* to prove the universal statements and to deduce from the universal statements, and the *logical rules* to prove the existential statements and to deduce from existential statements.

– For every type T,
  – a *proof* of an universal statement "for every x of type T, $A[x]$" is a proof π of $A[a]$ where a is *a generic object* of type T in π, i.e. is an *uniform (parametric) proof* of $A[t]$ for each object t of T;
  – a *deduction* from an universal statement "for every x of type T, $A[x]$" starts with the application of "for every x of type T, $A[x]$" to an object

t of type T and continues with a deduction from the statement $A[t]$ (in particular, if t is an object of type T, the conclusion $A[t]$ may be immediately deduced from the universal statement "for every x of type T, $A[x]$").

- For every type T,
  - a *proof* of an existential statement "for some x of type T, $A[x]$" is a proof of $A[a]$ where a is an object of type T;
  - a *deduction* from an existential statement "for some x of type T, $A[x]$" is a deduction π from $A[a]$ where a is *a generic object* of type T; in particular, a deduction π of a proposition C from "for some x of type T, $A[x]$" is a deduction of C from $A[a]$ where a is a generic object in the deduction (and so a does not occur in C), i.e. is a uniform (parametric) deduction from $A[a]$ in the context C for each object t of type T.

It is immediate that:

- all the above rules on the proofs of the universal statements and the existential statements, and on the deductions from the universal statements and the existential statements, are *logical rules* since they hold for every type, i.e. they do not depend on a given particular type;
- the logical rule on proofs of the universal statements is linked with the notion of *generic object inside a proof*;
- the logical rule on the deductions from the existential statements is linked with the notion of *generic object inside a deduction*.

Therefore, the logical rule on the proofs of the universal statements allows to get – when an universal statement "for every x of type T, $A[x]$" is proved by means of this rule – to get in an uniform way, for each given object b of type T, a proof of the statement $A[b]$ which does not depend on the knowledge of the specific object b; and the logical rule on the deductions from existential statements allows to get – when we have a deduction from "for some x of type T, $A[x]$" obtained by means of this rule – in an uniform way, for each given object b of type T, a deduction from the statement $A[b]$ which does not depend on the knowledge of the specific object b.

1.4 Other rules on the proofs of the universal statements and on the deductions from existential statements

There are *other ways* to prove universal statements and to deduce from existential statements, and these ways are valid only for particular types and are not valid for

every type, or are depending on the specific objects of the types, so that are not *logical ways*, are not *logical rules*: indeed, a rule on proofs or on deductions may be considered as a *logical rule* when it may be applied to any type and it does not depend on the specific objects of types.

These non-logical rules to prove universal statements are linked with another view of the universal and existential statements, i.e. with the following very common view of these statements (a view used in the semantic treatment of the universal statements and existential statements):

- an universal statement "for every x of type T, $A[x]$" is considered as the *conjunction* of all the instances $A[a]$ where a is an object of type T, and so the length of this conjunction *depends on* the type T;
- an existential statement "for some x of type T, $A[x]$" is considered as a *disjunction* of all the instances $A[a]$ where a is an object of type T, and so the length of this disjunction *depends* on the type T.

Therefore, according to this view of the universal and existential statements, when we know a type T:

- if T is finite, each universal statement "for every x of type T, $A[x]$" is considered as a *finite conjunction*, and each existential statement "for some x of type T , $A[x]$" is considered as a *finite disjunction*; in particular, when the elements of T are n objects in a given order $a_1, \ldots, a_n$
  - each universal statement "for every x of type T, $A[x]$" is considered as the conjunction $A[a_1] \wedge \ldots \wedge A[a_n]$,
  - each existential statement "for some x of type T, $A[x]$" is considered as the disjunction $A[a_1] \vee \ldots \vee A[a_n]$;
- if T is infinite, each universal statement "for every x of type T, $A[x]$" is considered as an idealized *infinite conjunction*, and each existential statement "for some x of type T, $A[x]$" is considered as an idealized *infinite disjunction*; in particular, if T is denumerable and the elements of T are in the order $a_1, \ldots, a_n, \ldots$,
  - each universal statement "for every x of type T, $A[x]$" is considered as the idealized infinite conjunction $A[a_1] \wedge \ldots \wedge A[a_n] \wedge \ldots$,
  - each existential statement "for some x of type T, $A[x]$" is considered as the idealized infinite disjunction $A[a_1] \vee \ldots \vee A[a_n] \vee \ldots$.

Remark that infinite conjunctions and infinite disjunctions are idealized objects and cannot occur inside a proof or inside a deduction (since every proof and every deduction must be a finite object).

This view of the universal and existential statements is compatible, of course, with the logical rules concerning the proofs of these statements and

the deductions from these statements; but this view allows to accept some other *non-logical rules* to prove the universal statements and to deduce from the existential statements.

A first non-logical rule to prove the universal statements, and a first non-logical rule to deduce from the existential statements, may be applied only on the universal or the existential statements on *finite* types. These rules may be formulated as follows: if T is a *finite* type with n objects,

- we may prove an universal statement "for every x of type T, $A[x]$" by proving $A[a]$ for each object of T, i.e. by getting n proofs, i.e. by a n-ary inference rule;
- we may deduce from an existential statement "for some x of type T, $A[x]$" by deducing from $A[a]$ for each object a of T, i.e. by producing n deductions, i.e. by a n-ary inference rule.

This rule to prove the universal statements, and this rule to deduce from the existential statements, may be accepted only when these statement are on a finite type, since otherwise (i.e. in the case of an infinite type T) the proof would contain *infinite* proofs, one proof for each object of T: and no proof may be infinite. Therefore, this way to prove the universal statements and this way to deduce from existential statements are not logical rules since they are restricted to finite types..

Another non-logical rule to prove the universal statements and another non-logical rule to deduce from the existential statements are the following rules which are valid for finite or infinite types (i.e. for every type) but are depending on the specific objects of the types. These rules may be formulated as follows: given a type T,

- we may prove an universal statement "for every x of type T, $A[x]$" in a given context by showing a procedure which allows, given any particular object a of T, to obtain a proof π_a of $A[a]$ in the same context and each π_a depends on the particular object a;
- we may deduce from an existential statement "for some x of type T, $A[x]$" by showing a procedure which allows, given any particular object a of T, to obtain a deduction π_a from $A[a]$ in the same context and each π_a depends on the particular object a.

Remark that these rules to prove the universal statements, and to deduce from the existential statements, are not logical rules since the procedure which gives a proof – or a deduction – for each object a of a type T may depend on the type T and on the specific object a:, whereas the logical rule to prove the universal statements and the logical rule to deduce from the existential statements are exactly the cases when the procedure does not depend on the type and does not depend on the specific object, i.e. the cases where the procedure is uniform.

I show two very interesting examples of these not logical rules to prove universal statements, when the universal statements are on the type of the natural number: the first example is the binary rule called *induction rule*, and the second example is the rule called *constructive ω-rule*.

1. The *induction rule* is the following *arithmetical rule* to prove universal statements on natural numbers: a statement "for every natural number a, $A[a]$" in a context Γ is proved form the following two premisses
   - the first premise is a proof π of $A(0)$ in the context Γ,
   - the second premise is a proof ψ of $A(a) \to A(a+1)$ in the context Γ where a is a generic natural number.

   Indeed, the premisses of the induction rule produce a procedure giving a proof π_n of $A(n)$ in the context Γ, for each given natural number n:
   - the proof π_0 of $A[0]$ in the context Γ is the proof π stated in the first premise,
   - the proof π_1 of $A(1)$ in the context Γ is obtained from the proof π_0 of $A[0]$ in the context Γ and the proof ψ of $A(0) \to A(1)$ (i.e. the proof ψ stated in the second premise, by replacing the generic object a by the natural number 0),
   - ...
   - the proof π_{n+1} of $A[n+1]$ in the context Γ is obtained from the proof π_n of $A[n]$ in the context Γ and the proof ψ of $A(n) \to A(n+1)$ (i.e. the proof ψ stated in the second premise, by replacing the generic object a by the natural number n),
   - ...
2. The *constructive ω-rule* is the following *arithmetical rule* to prove universal statements on natural numbers: a statement "for every natural number x, $A[x]$" in a context Γ is proved when there is a computable function giving, for each natural number n, the proof π_n of $A[n]$ in the context Γ. Induction rule is, indeed, a particular case of the constructive ω-rule.

2 The operators τ and ϵ: a proof-theoretical way to represent universal statements and existential statements

Hilbert's operators τ and ϵ allow to represent the universal statements and the existential statements in a different way w.r. to the more usual representation of these statements based on quantifiers $\forall$ and $\exists$ (i.e. the representation of "for every

x of type T, $A[x]$" by $\forall x A[x]$ and the representation of "for some x of type T, $A[x]$" by $\exists x A[x]$, when x is a variable of type T).

The operator τ has been introduced by Hilbert – together with Paul Bernays – before the operator ϵ. Indeed:

- the first public exposition of the operator τ is in the Hilbert's talk *Die logische Grundlagen der Mathematik* (1922, published in 1923),
- the first public exposition of the operator ϵ in used in the Hilbert's talk *Über das Unendliche* (1925, published in 1926) and the operator ϵ in present in the following Hilbert's talks devoted to the foundations of mathematics and in the book *Grundlagen der Mathematik* 1934 by Hilbert and Bernays.

Hilbert introduced the operator τ in the talk *Die logischen Grundlagen der Mathematik*, and the operator ϵ in the following talks, with the aim to express the universal statements and the existential statements of mathematical analysis (since these operators are used in the formalization of mathematical analysis as a first step towards the proof of the consistency of this discipline), but also to express the universal statements and the existential statements in the ordinary language (since he explains this operator by using examples of statements coming from the ordinary language). Thus, we may say that Hilbert aimed to represent, by τ-operator or by ϵ-operator, the universal statements and the existential statements on every type T.

I will show (in 2.1 and 2.2) that the new representation of the universal statements and the existential statements, proposed by Hilbert by means of the operators τ and ϵ, is very close to the logical rules concerning the proofs of these statements and the deductions from these statements. Therefore, we may say that the representation of the universal statements and the existential statements by means of Hilbert's operators τ and ϵ is a *proof-theoretical representation* of the universal statements and the existential statements.

Usually, one says that the motivations of David Hilbert for introducing the operators τ and ϵ are related to the formulation of the principle of choice (indeed, the operator ϵ is close to a choice operator), to the distinction between real elements (the propositions not containing the operators ϵ and τ) and ideal elements (the propositions containing the operators ϵ or τ), or to defend the impredicative definitions against the criticisms of predicativism and intuitionism.

In 2.3 I will emphasize that all these motivations become very clear, when the operators τ and ϵ are considered as tools for a proof-theoretical representation of the universal and the existential statements.

2.1 Representation of the universal statements and existential statements by means of the operator τ

Hilbert's operator τ gives, for every type T and every formula $A[x]$ (where x is a variable of type T), a τ-term $\tau x A[x]$.

Hilbert's operator τ is defined by the τ-axiom

$$A[\tau x A[x]] \to A[x]$$

for every type T and every variable x of type T, together with the substitution rule which allows to replace in the τ-axiom the free variable x of type T by an arbitrary object t of the type T, i.e. to obtain for every object t of type T

$$A[\tau x A[x]] \to A[t]$$

Hilbert proposed the following representation of the universal and existential statements by means of the operator τ:

- to represent each universal statement "for every x of type T, $A[x]$" by $A[\tau x A[x]]$
- to represent each existential statement "for some x of type T, $A[x]$" by $A[\tau x \neg A[x]]$

Why this representation?

I propose a *proof-theoretical answer* to this question, i.e. an answer linked to the logical rules to prove the universal and existential statements and to deduce fron these statements.

Firstly, I say that by the τ-axiom and the substitution rule, when x is a variable of type T:

- each τ-term $\tau x A[x]$ is exactly the *generic element a of type T inside some proof of* $A[a]$,
- each τ-term $\tau x \neg A[x]$ is exactly the *generic element a of type T inside some deduction from* $A[a]$.

Indeed, let x be a variable of type T:

1. when $\tau x A[x]$ denotes the *generic element a of type T inside some proof* π *of* $A[a]$, then

   $$A[\tau x A[x]] \to A[t]$$

   is true for every object t of type T because:

- firstly, when $\tau x A[x]$ denotes the *generic element a of type T inside some proof of $A[a]$*, if we prove $A[\tau x A[x]]$ then we prove the universal statement "for every x of type T, $A[x]$" (by the logical rule concerning the proofs of the universal statements) and so for every object t of type T we may prove the proposition $A[t]$ (by the logical rule concerning the deductions from the universal statements),
- finally, by replacing in the previous sentence the verb "we prove" by "holds", we get: if $A[\tau x A[x]]$ holds, then the universal statement "for every x of type T, $A[x]$" holds and so for every object t of type T the proposition $A[t]$ holds, i.e. $A[\tau x A[x]] \to A[t]$ holds;

2. when $A[\tau x A[x]] \to A[x]$ is true, then $\tau x A[x]$ is the *generic element a of type T inside some proof π of $A[a]$*: if π is a proof of $A[\tau x A[x]]$, since by the hypothesis $A[\tau x A[x]] \to A[x]$ is true, we may replace in π the term $\tau x A[x]$ by a free variable of type T not occurring in the context of the proof π, and this means that $\tau x A[x]$ is a generic object inside the proof π of $A[\tau x A[x]]$;
3. when $\tau x \neg A[x]$ denotes the *generic element a of type T inside some deduction π from $A[a]$*, then

$$A[t] \to A[\tau x \neg A[x]] \text{ i.e. } \neg A[\tau x \neg A[x]] \to \neg A[t]$$

   is true for every object t of type T because:
   - firstly, when $\tau x \neg A[x]$ denotes the *generic element a of type T inside a deduction from $A[a]$* and t is an object of type T, from a proof of $A[t]$ we may prove the existential statement "for some x of type T, $A[x]$" (by the logical rule concerning the proofs of the existential statements) and then we may prove $A[\tau x \neg A[x]]$ (by the logical rule concerning the deductions from the existential statements),
   - finally, by replacing in the previous sentence the verb "we prove" by "holds", we get: when t is an object of type T, if $A[t]$ holds then $A[\tau x \neg A[x]]$ holds, i.e. $A[t] \to A[\tau x \neg A[x]]$ holds;
4. when $A[x] \to A[\tau x \neg A[x]]$ is true, then $\tau x \neg A[x]$ is the *generic element a of type T inside a deduction π from $A[a]$*: if π is a deduction from $A[\tau x A[x]]$, since by the hypothesis $A[x] \to A[\tau x \neg A[x]]$ is true, we may replace in π the term $\tau x \neg A[x]$ by a free variable x of type T not occurring in the context of the deduction π, and this means that that $\tau x \neg A[x]$ is a generic object inside the deduction π from $A[\tau x A[x]]$.

For every formula $A[x]$ with x variable of type T, the formula

$$A[\tau x A[x]]$$

may be considered – as proposed by Hilbert – a representation of the universal statement "for every x of type T, $A[x]$" , since the premise of the logical rule to prove the statement "for every x of type T, $A[x]$" is a premise to prove $A[\tau x A[x]]$ and viceversa, and the premise of the logical rule to deduce from the statement "for every x of type T, $A[x]$" is a premise to obtain a deduction from $A[\tau x A[x]]$ and viceversa:

- when a is a generic element of type T in a proof π of $A[a]$ (i.e. when we may prove the statement "for every x of type T, $A[x]$"), we prove $A[\tau x A[x]]$ and viceversa, because $\tau x A[x]$ is the generic element a of type T in a proof of $A[a]$;
- when from the statement "for every x of type T, $A[x]$" we deduce $A[t]$ for some object t of type T, we may deduce $A[t]$ for some object t of type T also from $A[\tau x A[x]]$ by the τ-axiom and substitution rule, and viceversa.

When an universal statement "for every x of type T, $A[x]$" is expressed as $A[\tau x A[x]]$, then each proof of this universal statement is *uniform*! So, Hilbert's τ may be considered the first in the last century to emphasize the importance and the role of the uniform proofs.

Analogously, for every formula $A[x]$ with x variable of type T , the formula

$$A[\tau x \neg A[x]]$$

may be considered – as proposed by Hilbert – a representation of the existential statement "for some x of type T, $A[x]$" , since the premise of the logical rule to prove the statement "for some x of type T, $A[x]$" is a premise to prove $A[\tau x \neg A[x]]$ and viceversa, and the premise of the logical rule to deduce from the statement "for some x of type T, $A[x]$" is a premise to deduce from $A[\tau x \neg A[x]]$ and viceversa:

- when t is an object of type T and π in a proof π of $A[t]$ (i.e. when we may prove the statement "for some x of type T, $A[x]$"), we prove $A[\tau x \neg A[x]]$ by the τ-axiom and substitution rule, and viceversa;
- when a is a generic element of type T in a deduction π from $A[a]$ (i.e. when we have a deduction from the statement "for every x of type T, $A[x]$"), we have a deduction from $A[\tau x A[x]]$ and viceversa, because $\tau x A[x]$ is the generic element a of type T in a deduction of $A[a]$.

2.2 Representation of the universal statements and existential statements by means of the operator ϵ

Hilbert's operator ϵ gives, for every type T and every formula $A[x]$ (where x is a variable of type T), a ϵ-term $\epsilon x A[x]$.

Hilbert's operator ϵ is defined by the ϵ-axiom

$$A[x] \to A[\epsilon x A[x]]$$

for every type T and every variable x of type T, together with the substitution rule which allows to replace in the ϵ-axiom the free variable x of type T by an arbitrary object t of the type T, i.e. to obtain for every object t of type T

$$A[x] \to A[\epsilon x A[x]]$$

Hilbert proposed the following representation of the universal and existential statements by means of the operator ϵ (dual to the one given by means of the operator τ):

- to represent each universal statement "for every x of type T, $A[x]$" by $A[\epsilon \neg A[x]]$
- to represent each existential statement "for some x of type T, $A[x]$" by $A[\epsilon x A[x]]$

Why?

I propose a *proof-theoretical answer* to this question, which is dual of the analogous question about the representation by means of τ-operator.

Firstly, I say that by the ϵ-axiom and the substitution rule, when x is a variable of type T:

- each ϵ-term $\epsilon x A[x]$ is exactly the *generic element a of type T inside some deduction from $A[a]$*,
- each ϵ-term $\epsilon x \neg A[x]$ is exactly the *generic element a of type T inside some proof of $A[a]$*.

Indeed, let x be a variable of type T:

1. when $\epsilon x A[x]$ denotes the *generic element a of type T inside some deduction π from $A[a]$*, then

   $$A[t] \to A[\epsilon x A[x]]$$

   is true for every object t of type T because:
   - firstly, when t is an object of type T, if we prove $A[t]$ then we prove the existential statement "for some x of type T, $A[x]$" (by the logical rule concerning the proofs of the existential statements) and we may prove $A[\epsilon x A[x]]$ (by the logical rule concerning the deductions from the

existential statements, because $\epsilon x A[x]$ denotes the *generic element a of type T inside a deduction from* $A[a]$),

- finally, by replacing in the previous sentence the verb "we prove" by "holds", we get: when t is an object of type T, if $A[t]$ holds, then the existential statement "for some x of type T, $A[x]$" holds and so $A[\epsilon x A[x]]$ holds, i.e. $A[t] \to A[\epsilon x A[x]]$ holds;

2. when $A[x] \to A[\epsilon x A[x]]$ is true, then $A[\epsilon x A[x]]$ is the *generic element a of type T inside some deduction* π *from* $A[a]$: if π is a deduction from $A[\epsilon x A[x]]$, since by the hypothesis $A[x] \to A[\epsilon x A[x]]$ is true, we may replace in π the term $\epsilon x A[x]$ by a free variable of type T not occurring in the context of the proof π, and this means that $\epsilon x A[x]$ is a generic object inside the deduction π from $A[\epsilon x A[x]]$;
3. when $\epsilon x \neg A[x]$ denotes the *generic element a of type T inside some proof* π *of* $A[a]$, then

$$A[\epsilon x \neg A[x]] \to A[t] \text{ i.e. } \neg A[t] \to \neg A[\epsilon x \neg A[x]]$$

is true, because:

- firstly, when $\epsilon x \neg A[x]$ denotes the *generic element a of type T inside a proof of* $A[a]$, from a proof of $A[\epsilon x \neg A[x]]$ we get a proof of the universal statement "for every x of type T, $A[x]$" (by the logical rule concerning the proofs of the universal statements), and so for every object t of type T we prove $A[t]$ (by the logical rule concerning the deductions from the universal statements),
- finally, by replacing in the previous sentence the verb "we prove" by "holds", we get: if $A[\epsilon x \neg A[x]]$ holds, then $A[t]$ holds (where t is an object of type T), i.e. $A[\epsilon x \neg A[x]] \to A[t]$ holds;

4. when $A[\epsilon x \neg A[x]] \to A[x]$ is true, then $\epsilon x \neg A[x]$ is the *generic element a of type T inside a proof* π *of* $A[a]$: if π is a proof of $A[\epsilon x \neg A[x]]$, since by the hypothesis $A[\epsilon x \neg A[x]] \to A[x]$ is true, we may replace in π the term $\epsilon x \neg A[x]$ by a free variable x of type T not occurring in the context of the proof π, and this means that $\epsilon x \neg A[x]$ is a generic object inside the proof π of $A[\epsilon x A[x]]$.

For every formula $A[x]$ with x variable of type T , the formula

$$A[\epsilon x A[x]]$$

may be considered – as proposed by Hilbert – a representation of the existential statement "for some x of type T, $A[x]$" , since the premise of the logical rule to prove the statement "for some x of type T, $A[x]$" is also a premise to prove

$A[\epsilon x A[x]]$ and viceversa, and the premise of the logical rule to obtain a deduction from the statement "for some x of type T, $A[x]$" is also a premise to obtain a deduction from $A[\epsilon x A[x]]$ and viceversa:

- when t is an object of type T and π is a proof of $A[t]$ (i.e. when we may prove the statement "for some x of type T, $A[x]$"), then we prove $A[\epsilon x A[x]]$ by the ϵ-axiom and substitution rule, and viceversa;
- when a is a generic element of type T in a deduction π from $A[a]$ (i.e. when we have a deduction from the statement "for some x of type T, $A[x]$"), we have a deduction from $A[\epsilon x A[x]]$ because $\epsilon x A[x]$ is the generic element a of type T in a deduction from $A[a]$, and viceversa.

Analogously, for every formula $A[x]$ with x variable of type T, the formula

$$A[\epsilon x \neg A[x]]$$

may be considered – as proposed by Hilbert – a representation of the universal statement "for every x of type T, $A[x]$" , since the premise of the logical rule to prove the statement "for every x of type T, $A[x]$" is also a premise to prove $A[\epsilon x \neg A[x]]$ and viceversa, and the premise of the logical rule to obtain a deduction from the statement "for every x of type T, $A[x]$" is also a premise to obtain a deduction from $A[\epsilon x \neg A[x]]$ and viceversa:

- when a is a generic object of type T in a proof π of $A[a]$ (i.e. when we may prove the statement "for every x of type T, $A[x]$"), we prove $A[\epsilon x \neg A[x]]$ because $\epsilon x \neg A[x]$ is the generic element a of type T in a proof of $A[a]$, and viceversa;
- when from the statement "for every x of type T, $A[x]$" we deduce $A[t]$ for some object t of type T, we may deduce $A[t]$ also from $A[\epsilon x \neg A[x]]$ by the ϵ-axiom and substitution rule, and viceversa.

2.3 The operators τ and ϵ: impredicative definitions, choice principle, ideal elements

2.3.1 Impredicative definitions

An *impredicative definition* of an object c is a definition of c where there is an universal statement or an existential statement on a type T and c is an object of the type T.

In the critics of the impredicative definitions (e.g. in the critics made by Brouwer, Poincaré, Weyl,...), the usual objection against the use of such a kind

of definition in mathematics is the fact that in the *definiens* of c there is already the *definiendum* c.... but this objection holds when the universal statements on the type T are considered as *conjunctions* on all the elements of T and the existential statements as *disjunctions* on all the elements of T.

Instead, when an universal statement "for every x of type T, $A[x]$" is considered as $A[\tau x A[x]]$, and an existential statement "for some x of type T, $A[x]$" is considered as $A[\epsilon x A[x]]$, there is *no ground* for the usual criticisms: in this view, in the universal statements on a type T and in the existential statements of a type T we do not consider the elements of the type T.

2.3.2 Choice principle

The *choice principle* – in its general form – says that, given any collection of sets such that each set contains at least one object, it is possible to make a *selection* of exactly one object from each set and the results of this selection are the elements of a set.

Now, sets may be represented by formulas (for example, a set X may be represented by a formula $A[x]$ such that $A[t]$ holds if and only if t belongs to the set X) and a set represented by a formula $A[x]$ is non-empty when the existential statement "for some x of type T, $A[x]$" holds, i.e. by ϵ-axiom when $A\epsilon x A[x]$ holds (and by τ-axiom when $A[\tau x \neg A[x]]$ holds): so, if a non-empty set is represented by a formula $A[x]$, then $\epsilon x A[x]$ (or $A[\tau x \neg A[x]]$) is a selected element of this set.

So, by accepting ϵ-axiom or τ-axiom (i.e. when an existential statement "for some x of type T, $A[x]$" is considered as $A[\epsilon x A[x]]$ or is considered as $A[\tau x \neg A[x]]$), one accepts a general choice principle: the selected element of a non-empty set represented by the formula $A[x]$ is $\epsilon x A[x]$ or $\tau x \neg A[x]$.

But we have to remark that the operator ϵ *is not exactly a choice function*, since:

- firstly, the operator ϵ refers to formulas whereas choice function refers to sets,
- moreover, the operator ϵ is defined for each formula, whereas choice function is defined only for non-empty sets,
- finally, the object selected by the operator ϵ applied to a formula $A[x]$ is the generic object a inside some deduction from $A[a]$ which does not exists outside deductions.

2.3.3 Ideal elements

A term of the form $\tau x A[x]$ or $\epsilon x A[x]$ denotes a generic object of a given type *inside some proof or some deduction*, and nothing allows us to say that such *generic objects* exists outside proofs or outside deductions.

Moreover, these terms belong to the part of logic and mathematics which is not finitist and is called by Hilbert the *ideal part* of logic and mathematics: terms of the form $\tau x A[x]$ or $\epsilon x A[x]$ correspond to *ideal elements* of logic and mathematics, a topic discussed by Hilbert in *Über das Unendliche* (1926). A proof of the consistency of τ-axioms and ϵ-axioms would allow to say that the hypothesis of the existence of such generic objects is safe.

The *semantics* of these terms is very hard, and is very distant from Hilbert's approach to logic and foundations of mathematics. Indeed, let T be a type and x a variable of type T: it is difficult to have an idea of what is a denotation of $\tau x A[x]$ and $\epsilon x A[x]$ because of the following remarks:

- the denotation of $\tau x A[x]$ must be an object b such that, when $A[b]$ holds, then "for every x of type T, $A[x]$". If the type T is finite and "for every x of type T, $A[x]$" holds, $\tau x A[x]$ may be the last element checked in order to verify that every object of T has the property expressed by A; but what when T is not finite? and what when "for every x of type T, $A[x]$" does not hold? Hilbert proposed an example outside mathematics: if $A[x]$ is "x is a good man", then $\tau x A[x]$ is the man who is considered the worst man, since if this man *becomes* good then every man is good!
- the denotation of $\epsilon x A[x]$ must be an object b such that, when "for some x of type T, $A[x]$", then $A[b]$ holds. If the type T is finite and "for some x of type T, $A[x]$" holds, $\epsilon x A[x]$ may be the first element of T checked as an element enjoying the expressed by A. But what when T is not finite? and what when "for some x of type T, $A[x]$" does not hold? When the type T is infinite, a choice function would be able to give the value of the terms $\epsilon x A[x]$ such that "for some x of type T, $A[x]$" (i.e. when the property expressed by the formula is not empty), but not for the other ϵ-terms.

Remark that the real interest of Hilbert is to give an interpretation of each τ-term and each ϵ-term *inside a given proof*, when this proof ends with a formula which does not contain τ or ϵ, i.e. a formula without ideal elements.

3 Logical naturaleness and logical depth of τ-axiom and ϵ-axiom

In Hilbert's axiomatic approach, axioms on a concept are the *implicit definition* of the concept: so, τ-axiom and ϵ-axiom are implicit definition of the operators τ and ϵ.

I propose a logical analysis of τ-axiom and ϵ-axiom, i.e. a logical analysis of the operators τ and ϵ, by investigating what happens when in sequent calculus we prove the (more complex) identity $\forall x A[x] \vdash \forall x A[x]$ and $\exists x A[x] \vdash \exists x A[x]$ from the (more simple) identity $A[a] \vdash A[a]$.

The result of this logical analysis is:

- τ-axiom arises, in a very natural way, in some hidden steps of the derivation of $\forall x A[x] \vdash \forall x A[x]$ from $A[a] \vdash A[a]$,
- ϵ-axiom arises, in a very natural way, in some hidden steps of the derivation of $\exists x A[x] \vdash \exists x A[x]$ from $A[a] \vdash A[a]$.

This leads to say that τ-axiom and ϵ-axiom are logically natural and belong to a refinement of logic.

3.1 τ-axiom

I propose the following analysis of the τ-axiom $A[\tau x A[x]] \to A[a]$, i.e. $A[\tau x A[x]] \vdash A[a]$, by investigating what happens when we try to prove in sequent calculus the identity $\forall x A[x] \vdash \forall x A[x]$ from the identity $A[a] \vdash A[a]$.

In the proof of $\forall x A[x] \vdash \forall x A[x]$ from $A[a] \vdash A[a]$ in sequent calculus, we distinguish the following steps.

1. First step: we start with $A[a] \vdash A[a]$ where a is a generic object (a free variable) of a type T, with two occurrences of a (the left occurrence is in the formula in the left side of the sequent, the right occurrence is in the formula in the right side of the sequent).
2. Second step. The second step is the transition from $A[a] \vdash A[a]$ to $\forall x A[x] \vdash A[a]$ by using the left rule of the universal quantifier $\forall$. This step is split in two sub-steps:
   (a) firstly, we have to distinguish the left occurrence of a from the right occurrence of a, because we act on the left occurrence whereas we do not act on the right occurrence:
   - the left occurrence of a is considered as a particular object t of type T (in order to use the left rule of $\forall$)

– but we continue to consider the right occurrence of a as a generic object,

so that the left occurrence of a is considered as a particular object t of type T such that $A[t] \vdash A[a]$;

(b) then, we apply on the left side of the sequent the left rule of $\forall$ (which corresponds to the logical rule concerning the deductions from the universal statement, i.e. the rule of the use of the premise $\forall$), to conclude $\forall x A[x] \vdash A[a]$.

3. Final step. Since a does not occur free in the left side (and this means that we have a proof of $A[a]$ from the hypothesis $\forall x A[x]$) where a is a generic object of type T), then we use the right-rule of $\forall$ (the rule of proof of $\forall$, which corresponds to the logical rule concerning the proofs of universal statements) to conclude $\forall x A[x] \vdash \forall A[x]$.

If we stop at the step 2(a), i.e. if we stop before the introduction of quantifiers, the object t of type T such that $A[t] \vdash A[a]$ is just $\tau x A[x]$ and what we get is a τ-axiom : there is an object t such that, if $A[t]$ holds, then for every x of type T $A[x]$ holds.

3.2 ϵ-axiom

Now, I propose the following analysis of the ϵ-axiom $A[a] \to A[\epsilon x A[x]]$, i.e. $A[a] \vdash A[\epsilon x A[x]]$, by investigating what happens when we try to prove in sequent calculus the identity $\forall x A[x] \vdash \forall x A[x]$ from the identity $A[a] \vdash A[a]$

In the proof of $\exists x A[x] \vdash \exists x A[x]$ from $A[a] \vdash A[a]$ in sequent calculus, we distinguish the following steps.

1. First step: we start (as in the previous analysis of τ-axiom) with $A[a] \vdash A[a]$ where a is a generic object (a free variable) of a type T, with two occurrences of a (the left occurrence is in the formula in the left side of the sequent, the right occurrence is in the formula in the right side of the sequent).
2. Second step. The second step is the transition from $A[a] \vdash A[a]$ to $Aa \vdash \exists x A[x]$ by using the right rule of the existential quantifier $\exists$. This step is split in two sub-steps:
   (a) firstly, we distinguish left occurrence of a from the right occurrence of a, because we act on the right occurrence and we do not act on the left occurrence:
      – the right occurrence of a is considered as a particular object t of type T (in order to use the right rule of $\exists$),
      – whereas we continue to consider the left occurrence of a as a generic object,

so that the right occurrence of a is considered as a particular object t of type T such that $A[a] \vdash A[t]$;

(b) then, we apply on the right side of the sequent the right rule of $\exists$ (which corresponds to the logical rule concerning the proofs of the existential statements), to conclude $A[a] \vdash \exists x A[x]$.

3. Final step. Since a does not occur free in the right side (and this means that we have a proof of $\exists x A[x]$ of $A[a]$ where a is a generic object of type T), then by left rule of $\exists$ (the rule of use of existential statements, which corresponds to the logical rule concerning the deductions from existential statements) we conclude $\exists x A[x] \vdash \exists x A[x]$.

If we stop at the step 2(a), i.e. if we stop before the introduction of quantifiers, the object t of type T such that $A[a] \vdash A[t]$ holds is just $\epsilon x A[x]$ and what we get is an ϵ-axiom: there is an object t such that if for some x of type T $A[x]$ holds, then $A[t]$ holds.

Conclusion

I wish to mention other topics which I aim to consider in further papers devoted to the operators τ and ϵ in proof-theory, on the basis of the fact that these operators give a proof-theoretical representation of the universal and existential statements:

- Interdefinability of the operators τ and ϵ, and definability of the usual quantifiers $\forall$ and $\exists$ by means of the operators τ and ϵ;
- Quantifiers $\forall$ and $\exists$, and operators τ and ϵ, in sequent calculus
- Formalization of Arithmetics and Mathematical Analysis by means of the operators τ and/or ϵ
- Hilbert's procedure of elimination of ideal elements (τ-terms , ϵ-terms) in the proofs of formulas without ideal elements.

Tatiana Arrigoni

Truths in Contemporary Set Theory

1 Introductory remarks and contents

The aim of this paper is to sketch a brief overview of the contemporary debate in set theory about mathematical truth, and to give a critical appraisal of it. To this purpose I'll start by sketching an overall analysis of what is commonly and, perhaps, naively, meant by ascribing truth to sentences, focussing on the general conditions under which true sentences are usually said to be so. I'll then proceed by formulating the main thesis of the paper, i.e. the view that in contemporary set theory a process of "thinning out" truth has taken place in the last decades along with the mathematical development of the discipline: distinguished sentences of set theory are/continue to be regarded as true, and others are considered suitable candidates for new set theoretic truths, but when it comes to do so only minimal conditions for truth are invoked, with "old" assumptions concerning truth being bracketed and/or substituted with new, more "economical" ones. Of course, under these circumstances one may legitimately ask whether the pursuit of truth still makes sense in set theory today, or should better be dispensed with. A systematic discussion of this issue goes beyond the scope of this paper. I only observe that it seems not so easy to dispense with truth in contemporary set theory. The notion of truth has in fact displayed remarkable resilience in adapting to new results and unexpected mathematical scenarios, and continues to be a crucial term of reference for the practitioners. Any detailed analysis of this fact, however, must be postponed until to another occasion.

2 (Naive) truth inside and outside set theory: a brief analysis

What do we mean, implicitly and, perhaps naively, by claiming that sentences like e.g. "sets with the same elements are the same set" or "7+5 = 12", or, even, "the snow is white" are true?

As a first tentative answer to the question I suggest considering following conditions for truth, which for convenience I group under the three categories: ontological, epistemic and functional. Note the there is a relation of implication

Tatiana Arrigoni: Fondazione Bruno Kessler, Trento

DOI: 10.1515/9783110529494-003

among these conditions, the ones being mentioned first being sufficient (though not necessary) for the occurrence of the ones following them.

2.1 Ontological conditions

If we take a true sentence like "the snow is white" and examine why people are ready to agree that it is true, it seems straightforward to say that this happens because everybody is ready to acknowledge that the content expressed by it amounts to a state of affairs that holds objectively. By this I mean that it holds independently from the individual subject who happens to utter the sentence, more precisely from special conditions concerning him/her, e.g. distinguished cognitive features of him/her and/or the particular spatio-temporal coordinates at which he/she makes the utterance (or both). Below I'll examine what justifications for the objectivity of a state of affairs one may give. Here I just take for granted that, as a matter of fact, many are ready to connect truth with objectivity (as described above) and give some examples (in the negative) in support of my claim. If I wore pink glasses and I were on a snowy mountain in the Alps, I would see all white surfaces surrounding me as pink, but nobody (neither you nor myself) would be ready to regard the sentence "the snow is pink" as true, its content being clearly dependent upon a cognitive peculiarity of mine (the fact that I wear pink glasses). Analogously, since it is now January and I find in Italy I can legitimately say: "winter is a cold season". However, would you be ready to regard this sentence, as it stands, as true? I guess not. Were I in Argentina instead of Italy now, it would still be winter, but probably it would not be very cold. In fact, we are inclined not to regard the sentence "winter is a cold season" as true without further specifications. That its content holds, in fact, depends on the utterer's location. Moreover, a sentence whose content expresses a state of affairs that is recognized as holding independently from the subject who utters it, cannot but be regarded as mutually exclusive with its negation, i.e. the state of affairs expressed by the latter, even though possible per se, cannot be thought of as being actually the case, if the original sentence is true. The ontological condition implies bivalence.

2.2 Epistemic conditions

It seems also easy to agree that a sentence satisfying the conditions I called ontological, is likely to be endowed with epistemic features that make it especially compelling. As a result, all sound subjects presented with the sentence are expected to subscribe to it, i.e. the sentence is to be intersubjectively agreed upon.

Moreover its truth stands out as a property that admits no degrees and no fluctuations. It is a yes/no feature of the sentence: either it does possess it, or it doesn't. And if a sentence possesses it, its negation does not. We can make mistakes, of course, and regard as objectively holding states of affairs that later turn out not to be so, i.e., which is equivalent, take as true sentences that are later discovered to be false.[1] Truth-value revision, though, has no impact on the epistemic features of truth, it doesn't turn truth into a property that may change in time. If a sentence supposed to be true is later found to be false, it is regarded as such that it was never true after all. To state it again, truth is no property that may be only temporary there and/or modify in time – although we are ready to say this of our truth-ascriptions to sentences. Thus, as long as we think of a sentence as true, no matter whether we are right or not, we think of it as definitively so, and, at the same time we rule out that its negation can be true as well. Truth is to some extent "exclusive": no sentence can share it with its negation or sentences contradicting it.

2.3 Functional conditions

If 2.1 and 2.2 are the main features which are straightforwardly (albeit implicitly) ascribed to sentences in saying them true, further properties of truth follow. Let me specifically focus on one, which I call *functional* because it has to do not with truth *per se* but with the role true sentences are expected to play within the corpus of our knowledge concerning a subject. Being supposed to express states of affair that objectively hold, and being their epistemic status regarded as enduring in time and "exclusive", known true sentences are assumed to be inevitable starting points and constraints when it comes to enlarge our views and get more more and more information about a subject which remains partly obscure. One has to start from known true sentences about the subject, and not to contradict them, if one wants to know more about it.

3 (Non naive) Truth inside and outside set theory

The reason why I qualified above conditions for truth as possibly naive is because people may be ready to agree with them without any explicit justification for the possibility of the existence of true sentences. Are we fully legitimate in assuming that there are state of affairs that objectively hold (in the sense explained above),

1 Like the axiom of unrestricted comprehension, to make an example taken from the foundation of mathematics. See Jech (2003)

and to think that these can be expressed by true sentences? The most traditional way in which philosophers have given a positive answer to this question has been to claim the existence of a mind-independent reality, perfectly determined *per se*, knowable by the thinking subjects in a way that is faithful to how it is in fact, and correctly describable in words. Having used numbers from 1 through 3 for listing the above, possibly naive, features of true sentences, I now use number 0 for this new condition. Since it appears to be more fundamental than the others (which in fact can be implied by it), I call it "meta-condition" for truth or, more exactly, due to its content, realist "meta-condition" for truth.

3.1 The realist meta-condition for truth

State of affairs that objectively hold are realized in a mind-independent reality, determined *per se*, faithfully knowable by us, and describable in words via (true) sentences.

That true sentences possess eatures from 3.0 to 2.3, these being hierarchically ordered by a relation of implication according to their order (3.0 → 2.1 → 2.2 → 2.3), has been a widespread view in philosophy. It is usually called "correspondence theory of truth".

The most classical version of the theory is due to Aristotle.

> The fact of the being of a man carries with it the truth of the proposition that he is, and the implication is reciprocal: for if a man is, the proposition wherein we allege that he is, is true, and conversely, if the proposition wherein we allege that he is true, then he is. The true proposition, however, is in no way the cause of the being of the man, but the fact of the man's being does seem somehow to be the cause of the truth of the proposition, for the truth or falsity of the proposition depends on the fact of the man's being or not being. (*Categories* 14b14).

The correspondence theory of true, in slightly the same terms as formulated in the *Categories*, has been more than once invoked in set theory to account for the truth value of mathematical sentences. It is central to Platonism, as formulated by Gödel in the following quotation, where the truth/falsity of the statement expressing the Continuum Hypothesis is explained as depending on whether it reflects or not what is the case in the well-determined reality described by the axioms of set theory.[2]

2 The Continuum Hypthesis, formulated by Cantor, concerns the size of the set of all real numbers, i.e. the power set of the infinite set ω having all natural numbers as elements. The

> The set theoretical concept and theorems describe some well-determined reality, in which Cantor's conjecture must be either true or false. Hence its undecidability from the axioms assumed today can only mean that these axioms do not contain a complete description of that reality. (Gödel 1947/64, 476)

Also in more recent papers Platonism (and the related correspondence theory of truth) are invoked. In a recent contribution by Shelah, for example, the author appears reluctant to regard as a proper axiom for set theory any proposal advanced in the last decades for enlarging the system ZFC.[3] While doing so he repeatedly stresses that ZFC has a privileged status compared to more recent axiom candidates, and the point is made that proving a theorem in set theory means ultimately proving it in ZFC. Shelah's claim comes together with an endorsement: "I am a card carrying Platonist seeing before my eyes the universe of set". (Shelah (1993), p. 6). His "Platonist position" is described by the author in a later contribution as the view that "there is a unique universe of set about which we know no more than ZFC" (Shelah (2003), p. 215). It is clear from further considerations that ZFC is the only axiomatic system for sets that Shelah is ready to regard as true.[4] One has thus to conclude that implicit in both papers is the view according to which the truth of a statement about sets depends on and consists in reflecting what is the case in a unique, mind independent ("seen before one's eyes") universe of set.

4 The thesis of this paper *in nuce*

I am now in the position of stating the thesis of this paper more explicitly. The above quotations witness that there are set theorists according to whom distinguished set theoretic sentences have to be regarded as true because fulfilling all the conditions for truth listed above, i.e. the realist meta-condition, hence

conjecture is that the size of the power set of ω represents the smaller size for an infinite set above the one of ω itself. See Jech (2003) for mathematical details.

3 ZFC is the axiomatic system consisting of the Zermelo-Fraenkel axioms for set theory plus the Axiom of Choice. See Jech (2003).

4 Consider e.g. the discussion in Shelah (2003) about the axiom of constructibility, $V = L$, about which he asks: "Why the hell should it be true?" (p. 210). By this it is not meant that the axiom should be regarded as false, but only that it leads to one of the many possible set theories extending ZFC, all on a par as hypothetical scenarios about sets, no one deserving to be regarded as more faithful to "the unique universe of sets" than any other. About the axiom of constructibility $V = L$ and its implications, see Jech (2003).

the ontological, epistemic and functional conditions described in section 2. This suggests an obvious interpretation of my initial claim that truth has been "thinned out" in contemporary set theory: this has been done through a progressive disentanglement of truth from Platonism as endorsed by e.g. Gödel and Shelah. In fact, with very few exceptions[5], when truth is invoked in set theory today, it is no longer understood, at least overtly, as consisting in sentences' faithfully describing states of affairs holding in a mind-independent world of mathematical objects. How is it meant instead? In what follows I will discuss two possible alternatives to Platonism about truth in contemporary set theory, which I'll call *quasi-Platonism* and a *no-Platonism-at-all*.

Although not usually named in this way, *quasi-Platonism* is the position of those who literally describe truth in quite the same terms as Platonists but with one remarkable exception.[6] Quasi-Platonists take true sentences of set theory to express states of affairs that apparently hold objectively, are epistemically compelling and deserve to play the role of axioms. For them, however, sentences expressing objectively holding states of affairs need not be intended as mirroring what is the case in a mind-independent mathematical reality. Accounting for objectivity in a non Platonistic setting is the major philosophical challenge of quasi Platonism. I will return on it in a moment.

Not differently from quasi-Platonism, the *non-Platonist-at-all* account of set theoretic truth is often intertwined with the view that there are sentences of set theory that, although not yet so, do deserve to be regarded as (new) set theoretic truths. When it comes to explain what this exactly means, it is said that they should be intersubjectively agreed upon and regarded as starting points for enlarging our set theoretic views, exactly as candidates for new axioms of set theory. That this is so, though, is not explained by saying that these sentences express states of affairs that we know as holding objectively, i.e. independently from us. Nor are they said to have, at least *prima facie*, compelling epistemic features. Instead it is the mathematical developments in which they are involved that is supposed to endow them with a special status. This may be called "honorary truth", truth as a sign of mathematical merit. Note that honorarily true statements do clearly fulfil the functional conditions for truth mentioned above.

5 Shelah is one of them, and other set theorists may perhaps be ready to express similar views in conversation.

6 The expression "objectivity over objects" is also used for the position, which I call quasi-Platonism. The expression in to be found in Wang, to describe views of Gödel (see Wang (1996)), and has also been recently used for views about truth in set theory of a quasi-Platonist flavour in Hauser (2001).

It is clear, too, that, so described, they do not satisfy the realist meta-condition for truth. Less clear is whether they may satisfy the ontological and epistemic conditions.

To conclude this overview, let me add that there also seems to be a *nihilist* perspective about truth in contemporary set theory. This is endorsed by those who either think that no set theoretic sentence deserves to be regarded as true, hence refuse to ascribe truth to any and claim that any use of the term true in set theory is deceptive, or are extremely generous in ascribing truth to sentences, even to contradictory ones. This perspective has been by large affected by the discovery of sentences independent from the axiomatic system ZFC, and the work done on different models of it (or suitable extensions of it).[7] In this connection truth is taken to mean "true in a distinguished model of ZFC" and either ascribed to mutually contradictory sentences, each consistent with ZFC, or denied to any set theoretic formula except for the ZFC axioms and theorems, bivalence being regarded as an indispensable condition for truth. I won't focus on the nihilist perspective in this paper. In fact, on the one hand, this paper subscribes to the view that truth and bivalence are strictly related (the latter is intertwined with all the conditions describing truth considered above). Hence the expression "true in a model" is not taken here to be *per se* synonymous with "true" in set theory. On the other hand, my focus is on views according to which not only the axiomatic system ZFC but also suitable extensions of it may deserve to be regarded as true.

5 Quasi-Platonism

Quasi-platonism, in its most coherent form, is closely intertwined with a distinguished line of research in contemporary philosophy of mathematics, applying the resources of Husserl's phenomenology to account for objectivity and truth in set theory. Some of the overall assumptions which quasi-Platonism starts from, however, may be found in a plenty of contributions devoted to large cardinal axioms, exactly when it comes to justify the claim that the latter should be regarded as axioms of set theory as standard as the axioms of the system ZFC.[8]

7 A formula φ of the language of ZFC is said to be independent from ZFC if is both φ and its negation are consistent with it (i.e. there model of both "ZFC + φ" and "ZFC + $\neg\varphi$"). See Kunen (1980) for an introduction to independency proofs.

8 Large cardinals are cardinals whose existence cannot be proved in ZFC (since they imply the consistency of the latter). In contemporary set theory a distinction is made between small and large large cardinals, according to whether they are consistent with the axiom of constructibility

The first central assumption worth mentioning in this connection is that we have a clear and persuasive intuitive concept of set, the so called "iterative concept", implying a distinguished picture of the universe of all sets. According to the latter, the universe, called V, looks like a transfinite sequence of stages, and sentences stating the existence of large cardinals do describe features of V implicit in the iterative concept. Hence, so the argument, these sentences must be regarded as true for the iterative concept and the view of the universe implied by it. This is enough for many to say that they are simply true and deserve to be taken as new axioms for set theory.

This view was first formulated by Gödel in his 1947/64, where the point is made that the statements that (small) large cardinals exist must be regarded as axioms since they are suggested by the "iterative concept of set" (for the same reason they are also said to be intrinsically evident).

> [...] the axioms of set theory by no means form a system closed in itself but, quite on the contrary, the very concept of set on which they are based suggest their extension by new axioms which assert the existence of still further iteration of the operation "set of" [...] These axioms show clearly, not only that the axiomatic system of set theory as used today is incomplete, but also that it can be supplemented without arbitrariness by new axioms which only unfold the concept of set. (Gödel, 1983, p. 476)

How the existence of large cardinals in V can be viewed as implied by the iterative concept is briefly illustrated in Martin (1998), as follows:[9]

> According to the iterative concept sets are to be regarded as being formed in a transfinite sequence of stages and that the number of these stages is supposed to be "absolutely infinite". From this absolute infinity one derives the related principles of resemblance (there should be pairs of stages that are alike in any given respect) and reflection (there should be stages that look like the whole universe of sets in any given respect). From the reflection principles come precise reflection schemata in the formal language of set theory.(Martin (1998), p. 229)

To see why a principle like *resemblance* should lead to large cardinal axioms, just consider that the axiom stating the existence of an inaccessible cardinal amounts to the claim that there is a stage of the universe of all sets V_κ that resembles the relation in which the stage V_ω (i.e. the set of all natural numbers) stays to the stages indexed by finite natural numbers, i.e. it cannot be reached "from

$V = L$. See Kanamori (1994) and Jech (2003) for mathematical details, and Arrigoni (2007) for a philosophical analysis of large cardinals and their role in contemporary set theory.

9 I've chosen to quote this contribution among many because it is significantly included in a volume which entitled "Truth in Mathematics", exactly Martin (1998).

below" by applying ZFC operations for building new sets (like *Union* or *Power Set*) to stages V_λ for $\lambda < \kappa$. As to reflection, Martin's argument draws back to the *Reflection Principle* of Levy (1960) and Montague (1961), which states that for any ZFC-formula $\varphi(x_1, \ldots, x_n)$ and any ordinal β there is a limit $\alpha > \beta$ so that for any $x_1, \ldots, x_n$ in V_α, $\varphi(x_1, \ldots, x_n) \leftrightarrow \varphi^{V_\alpha}(x_1, \ldots, x_n)$.[10] This means that the set V_α, relative to the ZFC-language, is like the universe, i.e. every formula true in the universe, with parameters confined to sets in V_α, is already true in V_α. This seems to be a reasonable assumption. If the number of the stages of V has to be "absolutely infinite", i.e. it never ends, the universe of all sets cannot be exhausted by our knowledge.[11] Hence, whatever we know about the universe should be already true at some stage V_α. Indeed, reflecting the Reflection Principle to certain stages amounts to the postulation of large cardinal hypotheses, which, since implied by *reflection*, appear to be directly linked to the picture of the universe suggested by the iterative concept.[12]

Many have also insisted on the especially compelling character of the iterative concept.[13]

> […] the iterative conception of sets […] often strikes people as entirely natural, free from artificiality, not at all ad hoc and one they might perhaps have formulated themselves. (Boolos (1971), p. 489)

The same view is also implied by the claim that the iterative concept of set is *intuitive*:

> […] people do set theory by extensive appeal to their *intuition* and there is practically universal agreement on the correctness or incorrectness of the results thus obtained, as

10 $\varphi^{V_\alpha}(x_1, \ldots, x_n)$ is the relativization of $\varphi(x_1, \ldots, x_n)$ to V_α got by restricting the parameters of the formula to sets belonging to V_α.

11 The expression "absolutely infinite" was introduced by Cantor himself who distinguished the *Absolute* from the *Transfinite*. Where the former goes beyond any possibility of determination (and is ultimately identified with God), the latter is regarded as an increasable quantity, which displays different sizes, expressed by different (transfinite) cardinal numbers. See Cantor (1883), and Jané (1995) for a comprehensive analysis of Cantor's Absolute.

12 See Arrigoni (2007) for a more extensive illustration of reflection and large cardinals. Starting from a suggestion of Magidor, Bagaria has recently proposed to embrace not only reflection between stages of the universe but also reflection between structures within the universe. In this way one may derive most large cardinals known to now, which are thus said to be "natural" by Bagaria. See Bagaria (2005).

13 Many have contested, however, that the conception is partly only intelligible only due to our prior familiarity, got through purely mathematical work, with the principles/axioms it is supposed to shed light on. See e.g. Jané (2005).

> results about sets. The iterative concept of set is an *intuitive* concept and this concept has lead to no contradiction.(Wang (1996), p. 553)

Notice that the views expressed in these quotations do not amount to quasi-platonism. What is missing to them is an account that combines the equation "set theoretic truth = truth for the iterative concept" with a view of truth that, not satisfying the realist meta-condition, still does obey the ontological condition. That so an account is possible is not clear *prima facie*. In fact, on the one hand, the equation is *per se* compatible with full Platonism (just assume either that the universe of all sets described by the iterative concept is a "well determined reality" independent from us, or that that concept is in itself a well determined reality). On the other hand, the equation "set theoretic truth = truth for the iterative concept" may come together with the view that, far from satisfying the ontological condition, set theoretic truth does fail in doing so: if one embraces no Platonism with regard to either concepts or objects, one may see the iterative concept and the picture of the universe implied by it as men-made mental contents, and regard sentences true for the iterative concept and V as true of and for our "thoughts" (about sets and set structures). Under these circumstances they would express states of affairs that ultimately depend on us, holding because of us.

To see how quasi-Platonism may be formulated starting from the equation "set theoretic truth = truth for the iterative concept", recall that the ontological condition was described above as fulfilled if states of affairs expressed by sentences stand out as independent from cognitive or spatio-temporal features of the individuals that utter the latter. However, no mention was made of features that may be shared among many individuals. Thus, if one is able to argue that there is nothing involving any individual mathematician and his/her peculiarities in how "thoughts" like the iterative concept of set came to light – and in how we came to refer to it, explore it and make discourses about it – one may perhaps be in the position of arguing that, although expressing "thoughts" of ours about sets and set structures, the iterative concept, V, and what holds in it according to that concept, still can be seen as contents that objectively hold.

In fact in the literature one finds accounts according to which the iterative concept, and the related picture of the universe, have to be understood as powerful metaphors emerged through years of mathematical work, inspired by the results obtained all along in set theory and constrained by them. It was the shared effort of a community of research that has produced the iterative concept, and the picture of the universe "implied by it". This was possible once the overall features of a formal structure shaped by and emerged from purely mathematical work became well understood and people were able to convey (part of) its content

by an effective narrative (the iterative concept) and a pictorial illustration (the cumulative hierarchy V).[14]

The moral of the story is that, being no invention of any individuals, and being constrained by the mathematical results obtained all along by set theorists, the iterative concept, the universe V, and statements expressing what is true for them, are such that people can be intentionally related to them as to autonomous contents, independent from the individual subjects, no matter whether they ultimately originated in human cognition (long time mathematical work, paraphrase, and successful communication). Statements true for the iterative concept can thus be legitimately regarded as satisfying the ontological condition, hence the epistemic and the functional ones, without committing to Platonism.

Conclusions seemingly analogous to the above have been drawn by Gödel in 1947/64, where he uses the expression "mathematical intuition" for the act through which we are intentionally related to the iterative concept and the universe V as to autonomous contents.

> […] the question of the objective existence of the objects of mathematical intuition […] is not decisive for the problem under discussion here. The mere psychological fact of the existence of an intuition which is sufficiently clear to produce the axioms of set theory and an open series of extensions of them suffices to give meaning to the question of the truth or falsity of propositions like Cantor's Continuum Hypothesis.(Gödel (1983), pp. 484–485)

In 1947/64, however, Gödel is silent on how and from what the mathematical intuition he invokes, and its objects, may have originated.[15] These issues have been explicitly investigated by some contemporary authors. Following again a suggestion by Gödel, concepts and themes of Husserl's phenomenology are applied to give a systematic account of the constitution in human cognition of the objects of mathematical intentionality (concepts, structures, and objects in a proper sense), of the objectivity of the states of affairs that are known to be true for them and, hence, of mathematical truth. To enter into details of phenomenological accounts of objectivity and truth in mathematics goes beyond the scope of this paper. I refer to the relevant literature for suitable exemplifications.[16]

---

14 See my overview in Arrigoni (2011), pp. 337–60, 355 ff. Note also that, although referring back to Gödel 1947/64, accounts of set theoretical contents in terms of the iterative concept, first appeared in articles of the 1970's, whereas set theory was born at the end of the 19th century.

15 See Hauser (2006) for a very comprehensive phenomenological interpretation of Gödel's views.

16 See Tieszen (1989) for a general phenomenological account of mathematics, and Hauser (2001), Hauser (2006) for accounts focussed on set theory.

6 No-Platonism-at-all

Not differently from Quasi Platonism, *no-Platonism-at-all* starts from assumptions which are *per se* be compatible with alternative accounts of set theoretic truth, in particular with both Platonism and quasi-Platonism. In fact the original suggestion for no-Platonism-at-all stems from Gödel's papers. In his 1947/64 Gödel suggests that axioms that are very successful (extrinsically evident axioms) deserve to be regarded as true.

> […] even disregarding the intrinsic necessity of some new axiom, an even in the case it has no intrinsic necessity at all, a probable decision about its truth is possible also in another way, namely, inductively, by studying its success […] There might be axioms so abundant in their verifiable consequences, shedding so much light upon a whole field, and yielding such powerful methods for solving problems […] that no matter whether or not they are intrinsically necessary, they would have to be accepted. (Gödel (1983), p. 477)

Recognizing sentences of set theory as true since intertwined with successful mathematics, and suggesting that they should be adopted as axioms, lead to no-Platonism-at-all if one is ready to make two further claims. First one must contend that the truth of statements connected with successful mathematics is ultimately a matter of convention, although not an arbitrary one. It is grounded in the decision of taking distinguished sentences as constraints in enlarging our set theoretic knowledge, and reduces to the fact that a community of research converges on starting from them in enlarging its views, and on not revising them (if not provisionally, to see what consequences this may lead to).[17]

17 Steel's description (in his 2000) of what he thinks an axiom should be is worth quoting since he just points to the features I mention here.

> By *axiom* I shall mean: assumption to be adopted by all, as part of a broadest point of view. The "broadest point of view" proviso is meant to exclude from attention the temporary adoption of restrictive assumptions as a convenient device for avoiding irrelevant structure. $V = L$ is often assumed temporarily for such reasons by set theorists who do not believe it, just as "all functions are C^{∞}" is sometimes assumed by differential geometers who do not believe it. (Steel (2000), p. 422)

On the other hand, one must also be ready to acknowledge that there is no more in our decision of ascribing truth to "successful" sentences, and use them as axioms, than the fruitfulness of the mathematics they produced and/or are connected with. I.e. mathematical success and fruitfulness have not to be regarded as hinting at further properties of sentences, to be be invoked in the first place in ascribing truth to them and in accounting for the decision to do that. Instead

Hardly can one find no-Platonism-at-all formulated in these terms in contemporary set theory. Success is in fact often invoked as a hint of truth, but the latter is mostly intended as not resulting from success alone. In fact the truth of successful axioms is usually accounted for in a Platonist or quasi-Platonist way in set theory. This seems to be Gödel's case. This is also true of some authors who have focussed on the success of large cardinal axioms (and related principles) in more recent times. Often the point is made that having discovered the compatibility of some very successful, but not intuitively plausible, large cardinal axioms with principles inspired by the iterative concept and the absoluteness of the universe, has played a decisive role in regarding these axioms as legitimately true, as if success alone could hardly produce this result.

The case of the axiom of a measurable cardinal (MC) has been especially discussed in the literature. Although successful, it was felt to lack evidence in its original formulation (1930) and people were reluctant to take it as true.[18] The situation changed, however, since its connection with elementary embeddings was later discovered. Being truth preserving transformations of the universe V into a subclass resembling it, equipped with the standard $\in$ relation, elementary embeddings are seen as connected with *reflection*. Hence their existence is regarded as intrinsically evident. Today MC is usually presented as follows.

> Nowadays definitions of measurable cardinals and their generalizations are phrased in terms of *elementary embeddings* [...] A cardinal κ is measurable if and only if there is an elementary embedding j of V into a transitive target model M with κ being the least ordinal moved by j (the *critical point* of j). Progressively stronger large cardinal notions such as *supercompact cardinals* are obtained by demanding yet more resemblance between M and V. It turns out that through elementary embeddings a connection with the Absolute can be established lending support to Godel's contention that in the last analysis every axiom of infinity should be derived from the (extremely plausible) principle that V is undefinable. (Hauser, 2006, p. 536)[19]

truth is to be understood as a status that sentences come to be awarded by in the long run due to mathematical merits acknowledged to them.

18 Measurable cardinals were introduced by Ulam in 1930 as cardinals carrying a measure i.e. a function $\mu : \kappa \mapsto [0, 1]$ so that $\mu(\kappa) = 1, \mu x = 0$ for all $x \in \kappa$, and, for pairwise disjoint $\{X_n : n \in \omega\} \subseteq P(\kappa), \mu(\cup_n X_n) = S_n \mu(X_n)$. This is indeed equivalent to the existence of a non-principal κ-complete ultrafilter over κ, a property which generalizes a property of ω. Generalization seemed not enough, however, for seeing measurable cardinals plausible as they were first introduced. See Kanamori (1994) for technical details.

19 See also Bagaria (2005).

Analogous discourses are made about the axiom of Projective Determinacy (PD).[20] PD is described by Woodin as the true axiom for the structure $H(\omega_1)$.[21]

> There are natural questions about $H(\omega_1)$ which are not solvable form ZFC. However there are axioms for $H(\omega_1)$ which resolve this question, providing a theory as canonical as that of number theory, and which are clearly true. But the truth of these axioms became evident only after a great deal of work. For me, a remarkable aspect of this is that it demonstrates that the discovery of mathematical truth is not a purely formal endeavour. (Woodin (2001), p. 681.)

Woodin apparently refers here to the work done in set theory in order to both exploit the consequences of Projective Determinacy, and to establish the linkage existing between PD and large cardinal axioms. The latter are implicitly regarded as more suitable candidates for set theoretic truth than PD, since connected to the iterative concept and the absoluteness of V. This explains why the proof of PD from the existence of infinitely many Woodin's cardinal may be said to have decisively contributed to make the truth of PD fully evident.[22]

The change in the status of PD after being proved to depend on Woodin cardinals, is pointed out also by Hauser, with an interesting coda on Gödel.

> Unlike some large cardinal axioms, however, determinacy axioms are not evident by themselves, but are accepted mainly in view of their fruitful consequences. That raised the challenge to derive those consequences and even determinacy axioms themselves from large cardinals, a task that looked difficult given that large cardinals and determinacy seemingly bear no relation to each other. Nevertheless, in the end it has become clear that these two classes of axioms are really one class of axioms. Gödel might have interpreted this as a confirmation of his realistic views, for we would have no explanation for such an unexpected convergence unless these two kinds of axioms describe different aspects of one underlying reality. (Hauser (2006), p. 5)

---

20 PD is the axiom of projective determinacy, stating the determinacy of games having a projective set as payoff set. For $X \subseteq \mathbb{R}$, the infinite two-person *game* $G_X(A)$ associated to the payoff set $A \in {}^{\omega}X$ is defined as follows. There are two players, I and II; I initially chooses $x(0) \in X$, II chooses $x(1) \in X$, I chooses $x(2) \in X$. Let the resulting $x \in {}^{\omega}X$ be a play of the game. I wins if $x \in A$ otherwise II wins. A winning strategy for I is a function $s: \cup_{n\in\omega} {}^{2n}X \mapsto X$ so that , for $y \in {}^{\omega}X$ enumerating II's plays and $s * y$ denoting the play in which II plays y and I plays according to s, $s * y : y \in {}^{\omega}X \subseteq A$. A winning strategy for II is defined analogously. $G_X(A)$ is determined iff a player has a winning strategy. Both the projective hierarchy of sets of reals and PD are investigated in Moschovakis (1980).

21 $H(\omega_1)$ is the structure of sets of cardinality less than ω_1, the smallest uncountable cardinal.

22 The proof of PD from the existence of infinitely many Woodin cardinals is due to Martin-Steel (1989). Woodin cardinals are defined in terms of elementary embedding (see Kanamori (1994)).

In light of these quotations, it is no surprise that accounts of truth closer to no-Platonism-at-all are to be found in contemporary set theory when it comes to suggest that principles incompatible with large cardinals and/or PD do deserve to be regarded as new set theoretic truths. This is done e.g. within the program recently launched in Arrigoni, Friedman (2013), the so-called Hyperuniverse Program (HP). Within the HP one aims at enlarging the realm of set theoretic truths at an axiomatic level so as to arrive at solutions to questions independent from ZFC. To do so one selects among several candidate sentences, according to whether they fulfil criteria for truth stated in advance. More precisely, one adopts a multiverse perspective and consider all possible countable transitive models of the system ZFC, supposed to be an undeniable corpus of set theoretic truths (the *Hyperuniverse* is exactly the collection of all these models).[23] One then makes a selection among members of the Hyperuniverse preferring those who satisfy maximality properties stated in metamathematical terms: preferred members of the Hyperuniverse are models which neutralize the effects of possible manipulations of them (widening and lengthening) when it comes to prove consistency – hence they maximize consistency.[24] The members of the Hyperuniverse which are so so selected, are finally regarded as repositories for new set theoretic truths. In fact it is suggested that the statements they prove the consistency of (which include many solutions to independent questions) are assumed as new set theoretic axioms, as new truths to begin with to enlarge our views beyond ZFC. No matter if they turn out to be contradictory with (large) large cardinal axioms and PD (although not with their consistency).

It is worth stressing that within the Hyperuniverse Program, it is the procedure through which one arrives at selecting models and sentences holding in them that makes the latter suitable candidates to the role of new set theoretic axioms and new set theoretic truths, exactly its alleged reasonableness and its successful outcomes (after all, it leads to solutions to independent questions).[25]

> Sentences "true in V" are meant to be sentences that are or should be regarded by set-theorists as definitive, i.e., ultimate and not revisable. Within the Hyperuniverse Program two sorts of statements are regarded as being qualified for this status [being true]. The first are those set-theoretic statements that, due to the role that they play in the practice of set

23 See Arrigoni and Friedman (2013) for details and the choice of the multiverse as consisting of all countable tarnsitive models of ZFC as a suitable one to start with.

24 See Arrigoni and Friedman (2013) for details.

25 See quotation below from Arrigoni and Friedman (2013) for an account of the ZFC truth within the HP. As a result, within the Hyperuniverse Program, truth (or, better, truth beyond ZFC) stands out as the outcome of a reasonable convention.

> theory and, more generally, of mathematics, should not be contradicted by any further set-theoretic statement that aims to be itself accepted as ultimate and not revisable. Let us call these statements *de facto* set theoretic truths. The axioms of ZFC are examples of such truths. But secondly, within the Hyperuniverse Program, one is ready to regard as true in V statements that, beyond not contradicting de facto set-theoretic truth, obey a condition for truth explicitly established at the outset (i.e., they hold in all preferred universes of sets). Let us call these *de jure* set-theoretic truths. Note that, as intended by the Hyperuniverse Program, formulating de jure set theoretic truths is an autonomously regulated process. No external constraint is imposed while one is engaged in it, in particular there is no independently existing well-determined reality to which one must be faithful. Instead, in searching for de jure set-theoretic truths one is only expected to follow justiable procedures. It cannot be excluded at the outset that at some time the need will arise to modify the procedures adopted, in order to integrate them with other, equally reasonable procedures.(Arrigoni and Friedman, 2013, pp. 80–81)

It must be said, however, that this philosophical approach to truth, explicit in the first formulation of the program, has been partly abandoned in its most recent versions. In fact the focus now seems to be on the alleged *intrinsic* evidence of the metamathematical properties satisfied by the selected members of the Hyperuniverse. Since the latter may be described as maximality properties (the selected members of the hyperuniverse maximize consistency), they are understood as compatible with the iterative concept, implying the maximality or absoluteness of V.[26] Again one is faced with the implicit assumption that something more is needed to ascribe truth to sentences than a decision based on the success and the reasonableness of a selection procedure for them.

This is not the moral I want to draw here, though. In fact I contend that, no matter how it has evolved, the Hyperuniverse Program in its original formulation stands out as a relevant case study in the debate about truth in contemporary set theory. What it makes it interesting are not so much the concrete proposals for new set theoretic axioms that it leads to. In fact hardly principles contradicting large cardinal axioms are likely to find overall acceptance in today set theory. Not only the latter are clearly intrinsically evident, while the axiom proposals emerging form the HP seems not to be so, at least *prima facie*. Large cardinal axioms are extremely successful, too. Hence any decision to regard them as preferable to any sentence contradicting them has a solid rationale also from the perspective of no-Platonism-at-all. It is instead the way the Hyperuniverse Program addresses

26 See e.g. Antos et al. (2015). Notice that this claim may be debatable: looking at V as a structure (as an ontological domain) and focussing on it as model (a metamathematical device) may not be the same. Why an absolutely infinite structure should maximize consistency when employed as a model?

the quest for truth in set theory that makes it especially remarkable, and possibly a source of inspiration for new developments aimed at enlarging the corpus of set theoretic truths (beyond ZFC or, perhaps beyond ZFC + large cardinal axioms) in a multiverse scenario. One sees truth as primarily consisting in a function that sentences can have, and add that they can have it not only if they satisfy the ontological and the epistemic condition first. Instead one is ready to regard truth as the outcome of a decision which may require time and is grounded on mathematical developments and a reasonable agreement. Finally one is open towards an unusual scenario, with "thinner" truths possibly coming together with "fatter" ones, and consider how far this heuristic approach can lead him/her (beyond ZFC or, perhaps beyond ZFC + large cardinal axioms).

Marco Buzzoni

Gödel, Searle, and the Computational Theory of the (Other) Mind

Abstract: According to Sergio Galvan, some of the arguments offered by Lucas and Penrose are somewhat obscure or even logically invalid, but he accepts their fundamental idea that a human mind does not work as a computational machine. His main point is that there is a qualitative difference between the principles of the logic of provability and those of the logic of evidence and belief. To evaluate this suggestion, I shall first compare it with Searle's concept of "intentionality", and then introduce a distinction between two different senses of intentionality: a reflexive-transcendental sense and a positive (that is, historical empirical or formal logical) one. In the first of these senses, the nature of human reason is such that we have no idea how a real material system – or the corresponding formal one – could instantiate it. However, although this will turn out to be an important element of truth in Searle's and Galvan's conception, it does not exclude the opposite truth of Turing's functionalism: because intentionality, intuition, vision or insight – taken in their reflexive-transcendental sense – are simply invisible to the scientific eye, a man and a machine (or a robot) that is and one that is not endowed with intentionality are de facto indistinguishable from a strictly scientific point of view. For this reason, we might eventually be entitled, or even — by the practical precautionary principle – morally obliged, to attribute minds to machines.

1 Introduction

In an article about Gödel and the computational model of the mind, Sergio Galvan examined the implications of Gödel's incompleteness theorems for the philosophy of mind with special attention to the debate that was initiated by Lucas and Penrose (Galvan (2004)). According to Galvan, some of the arguments offered by these authors are somewhat obscure or even logically invalid, but he accepts their fundamental idea that a human mind does not work as a computational machine. His main point is that there is a qualitative difference between the principles of the logic of provability and those of the logic of evidence and belief. He develops this difference in the light of the considerations put forward in Gödel's Gibbs lecture "Some Basic Theorems on the Foundations of Mathematics

Marco Buzzoni: Macerata University

DOI: 10.1515/9783110529494-004

and Their Implications" (1951) and in a subsequent paper on the topic now reprinted in Gödel (1995), the third volume of *Gödel's Collected Works*.

The aim of this paper is to present and evaluate one particular aspect of Galvan's conclusions, namely that around which his notions of "vision", "intuition", and "understanding" revolve. In order to do this, it will be useful to compare this focal point of Galvan's reflections with Searle's Chinese Room thought experiment. For this reason, Sections 2 and 3 are devoted to a brief reconstruction of Galvan's and Searle's points of view, respectively. In Section 4, I shall try to show that Searle's concept of "intentionality" – but the same applies, *mutatis mutandis*, to Galvan's notions of vision, evidence, intuition, etc. – distinguishes the human mind from a (computational) machine only if it is understood in a much more radical sense than usual. For this purpose, I shall introduce the distinction between two different senses of intentionality: a reflexive-transcendental sense and a positive (that is, historical empirical or formal logical) one. In the first of these senses, the nature of human reason is such that we have no idea how a real material system – or the corresponding formal one – could instantiate it. However, paradoxically enough, although this insight is an important element of truth in Searle's and Galvan's conception, it also leads to the opposite truth contained in Turing's functionalism: from a strictly scientific point of view, intentionality, intuition, vision or insight – in their reflexive-transcendental sense – are simply invisible to the scientific eye. A man and a machine (or a robot) that is and one that is not endowed with intentionality are *de facto* indistinguishable from an empirical or from a logical point of view. The lacuna left open by any purely empirical evidence should be filled up by recourse to the practical precautionary principle, which, in particular circumstances or under particular conditions, might eventually morally oblige us to attribute minds to machines.

2 Galvan's reconstruction of Gödel's point of view

As already mentioned, Galvan's main concern is to test the soundness of the conclusions that, according to Luca and Penrose, can be drawn from Gödel's incompleteness theorems in the minds versus machines debate. According to Gödel, for every set of axioms which is sufficiently strong for arithmetic, if the system is consistent, there is what is now called a Gödelian sentence or formula that is true in ordinary arithmetic, but of which we cannot give a 'formal' proof 'in the system'. However, a human mind can recognise the truth of a Gödelian formula. According to Lucas and Penrose, this is sufficient to affirm that a human

mind, unlike any computational machine, is not bound by the rules of any formal system, and that there can be no axiomatization of it or of its thought processes. As Lucas says in a well-known passage (quoted by Galvan himself), "we can see that the Gödelian formula is true: any rational being could follow Gödel's argument, and convince himself that the Gödelian formula, although unprovable-in-the-system, was nonetheless – in fact, for that very reason – true." (Lucas (1961), pp. 113–115)

Briefly stated, the first argument discussed by Galvan is that mind's capacity for seeing that the Gödelian formula is true "in-the-system" makes it different in principle from any machine. Against this argument Putnam raised the seemingly fatal objection that, strictly speaking, the human mind can know that the Gödelian formula $G(T)$ is true only under the condition of knowing that the formal theory T is consistent, "which is unlikely if T is very complicated".[1]

It might be objected that "there are reasons which are known to the mind, but not to the machine for asserting the consistency of T" (Galvan (2004), p. 150). But, as Galvan rightly notes, these reasons can be in principle made treatable by mathematical methods, that is, they are accessible to the machine "to the extent that it is conceived dynamically, as a machine capable of acquiring always new information."[2]

---

1 Putnam (1960), p. 153. Galvan gives us a more precise logical formulation of Putnam's objection: cf. Galvan (2004), p. 152. Putnam's objection was taken up and developed, for example, by Shapiro (2003), pp. 25–26, and Raatikainen (2005), pp. 520–523 and 2015). McCall (1999), in a certain sense, may be regarded as an attempt to reverse this conclusion and to show that "[f]or the machine, the category 'not yet proven but true' does not exist." (McCall (1999), p. 531); however, in my opinion, Raatikainen (2002) has shown convincingly that McCall's argument is unsound. On McCall's paper, see also Gaifman (2000). On reflection, Putnam's objection finally develops a remark made by Turing against what he called "The Mathematical Objection". The objection is the following. In the imitation game, in the last analysis because of Gödel's incompleteness theorems, there will be some questions to which a digital computer "will either give a wrong answer, or fail to give an answer at all however much time is allowed for a reply." This should prove that there is "a disability of machines to which the human intellect is not subject." (Turing (1950), pp. 444–445) The first part of Turing's twofold reply is that which anticipates Putnam's objection: "although it is established that there are limitations to the powers of any particular machine, it has only been stated, without any sort of proof, that no such limitations apply to the human intellect." (Turing (1950), p. 445)

2 Galvan (2004), p. 150. This is also the point Turing made in his second reply to the already mentioned "Mathematical Objection". Turing says that "our superiority can only be felt on such an occasion in relation to the one machine over which we have scored our 'petty triumph'. There would be no question "of triumphing simultaneously over all machines. In short, then, there might be men cleverer than any given machine, but then again there might be other machines

For Galvan, the difficulties that the first argument must face led not only Lucas, but also Penrose to provide another argument in favour of the contention that the human mind is not a machine: whereas a mind knows, or at least believes that it is consistent, a machine does not know or believe it. As far as the machine in concerned, to know or to believe only means to "derive" (cf. Galvan (2004), pp. 150–151). Galvan takes up and improves Feferman's claim that this second argument contains the element of truth of the first, in the sense that "understanding is essential, and it is just this aspect of actual mathematical thought that machines cannot share with us." (Feferman (1995), par. 4.3; on this point, cf. also Salmon (2001), p. 100) It is precisely the working experience of mathematicians that distinguishes in principle between human and machine thinking (cf. Galvan (2004), p. 155).

In order to clarify this basic idea, Galvan compares the principles of the logic of evidence with those of the logic of provability. This comparison brings to light the difference between the functioning of the mind and the functioning of the machine and is Galvan's most distinctive contribution to this issue. For example, in contrast with provability theory, the principle of ω-completeness holds for the logic of evidence. According to Galvan's intensional reading, the possible formulation of this principle

$$\forall x \mathbb{E} A(x) \rightarrow \mathbb{E} \forall x A(x)$$

where $\mathbb{E}$ denotes the evidence operator, states that, if it is possible to make evident that every natural number n (mapping the numeral n) is A, then it is evident that all standard natural numbers are A (Galvan (2004), p. 158). Galvan admits that this is by no means sufficient to prove that there are truths which can be known by the human mind, but which are forever incapable of becoming known by a machine. However, he maintains that it means that the human mind follows different routes in the treatment of truths, because finitistic means are not sufficient to justify them (Galvan (2004), p. 163). This is, in his opinion, the real teaching of Lucas and Penrose. In this sense "evidence" (or intuition) "outdoes the performances that a machine is capable of, without meaning that the mind is capable of grasping the truth of sentences inaccessible to the machine." (Galvan (2004), p. 162)

How can this difference be evaluated and explained? Galvan's answer comes from the rereading of some pages which Gödel devoted to the consequences of his incompleteness theorems in his now well-known 1951 Gibbs lecture (Gödel (1995)). The most important consequence is that mathematics cannot

cleverer again, and so on." (Turing (1950), pp. 444–445.) On this point see also Feferman 1995, par. 4.5

be exhausted (or 'completed') by finitistic mathematics (in this sense, Hilbert's program in its original form was truly refuted by Gödel's theorems: see for example Raatikainen (2003)). According to Gödel, this admits of two explanations: "*Either mathematics is incompletable in this sense, that its evident axioms can never be comprised in a finite rule, that is to say, the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems.*" (Gödel (1995), p. 310; italics Gödel's. Cf. Galvan (2004), p. 163).

Because Gödel inclined to deny that there are "absolutely unsolvable diophantine problems", this disjunctive conclusion, to which for instance Shapiro (1998) had already called attention and which is also called "Gödel's dichotomy" (cf. Feferman (2006) and Feferman (2009), p. 209), is tantamount to cautiously affirm that the human mind "surpasses the powers of any finite machine", or, as Galvan prefers to say, that Gödel inclined to think that "the mind has resources that are not at the disposal of the machine" (Galvan (2004), p. 164).

However, as Galvan concedes, Gödel implicitly recognised that either alternative is theoretically possible.[3] In fact, the second one has been defended by many philosophers who share a naturalistic view of the mind, but it is easy to see that Turing was making the same point when he said that "it has only been stated, without any sort of proof, that no such limitations apply to the human intellect." (cf. footnote 1 of the present paper; see also Galvan (2004), p. 167 footnote, who, among others, referred to Benacerraf (1967), Dennett (1995), and Gaifman (2000)). According to this line of thought, the human mind is identical to a machine insofar as it may overcome the limitations that its work had uncovered using only finitistic means. If one admits that for any statement unprovable in a particular formal system, there are other formal systems in which the statement is provable, there is no reason to reject the same possibility for the concept of evidence, which might be regarded as a "contracted derivability" from the point of view of another theory (Galvan (2004), pp. 167–168).

Although either alternative is theoretically possible, according to Galvan some of Gödel's remarks weaken the last-mentioned one: "the thesis which identifies truth with derivability in a superordinate theory is legitimate only [...] under the condition that the superordinate theory is itself sound, that is, that it leads to consequences that are at least arithmetically true." To demand the soundness of Peano Arithmetic is tantamount to demand that Peano's axioms are justified. However, the fact that the axioms are a simple extension of Primitive

3 This has been recently reaffirmed by many authors: see, for instance, Feferman (2009), pp. 211–212; Raatikainen (2015), p. 45.

Recursive Arithmetic "is not sufficient to justify them". We need more substantive reasons, but of what kind are they?

According to the first alternative, to which Gödel and Galvan incline, they are "reasons of evidence": they consist in "the capacity to grasp [...] contents which are abstract and not finite, and into which we cannot gain insight through forms of finitistic evidence." (Galvan (2004), p. 164) In the case of the second alternative, they must be "inductive" or "selectionist reasons", since "reasons of evidence" are ruled out by the context (Galvan (2004), pp. 168–169).

In this way, Galvan tries to put the difference between minds and machines down to the difference between "reasons of evidence", on the one hand, and "inductive" or "selectionist reasons", on the other. And this way of putting the matter not only vindicates Gödel's idea that something is lost in translating the concept of proof, understood as that which provides evidence, into a merely formal or mechanistic concept, but can also be much more easily defended by such arguments as we find in Hume, Popper, or even Husserl.

Against inductivist and selectionist accounts, Galvan argues, for instance, that "induction presupposes kinds of abstract evidence, which are of intensional kind and resemble those of mathematical knowledge: for example, it is impossible to make sound inductions, without the power of grasping an adequate set of projectable predicates." (Galvan (2004, p. 169)) The same conclusion follows if induction is regarded as a method that makes it possible to pass from the particular to the general, from one particular case to the totality, to which it belongs. This is because in this case induction would coincide with the power of intuitively grasping general truths from the examination of a particular fact or object, that is, of the "structure of PRA and of the concrete and finite objects of its natural model".[4]

From this the more general conclusion is readily, though cautiously, drawn that "there are not sufficient grounds for regarding a computational model of the mind – based on the analogy between the mind and the functioning program (formal system) of the brain – as justified, even in the long run. On the contrary, the difficulties highlighted by a due consideration of the limitation theorems seem to point in different directions." (Galvan (2004), p. 172).

The arguments offered by Galvan are perhaps indirectly influenced by some Scholastic accounts of intuition, but Husserl's phenomenology has probably

4 Galvan (2004, p. 70). A similar, though not identical, point had already been made by Feferman (1995). Cf. for instance §40.3: "Mathematical Thought as it is actually produced is not mechanical; I agree with Penrose that in this respect, *understanding* is essential, and it is just this aspect of actual mathematical thought that machines cannot share with us." (italics in the original).

exerted the strongest influence on them, *through the mediation of Gödel himself*. In fact, as shown by Tieszen (1998), Husserl's ideas directly relevant to Gödel's comments in 1961 "are related to the fact that human cognition, including mathematical cognition, exhibits intentionality." (Tieszen (1998), p. 182). More precisely, both Husserl and Gödel maintain that there must be a kind of mathematical intuition or insight that accounts for our mathematical knowledge, that is, in Husserl's language, an intuition of "mathematical essences" (cf. Tieszen (1998), p. 189).

If we combine this with the obvious connexion between Gödel's incompleteness theorems and the Turing test, we are naturally led to a comparison between Galvan's conclusions and Searle's Chinese Room thought experiment. Already Gödel remarked that Turing's work gave an analysis of the concept of "mechanical procedure" or "computation procedure" that was equivalent with that of a "Turing machine" (Gödel (1986), p. 369. On this point, cf. for example Feferman (2009)). And what was for Turing a possible objection to be rebutted (cf. Turing (1950), p. 444) – or perhaps an objection already implicitly met by regarding the imitation game as a scientifically sound substitute for philosophical questions concerning "understanding" and "intuition" (for this interpretation, cf. Shapiro (1998)), was for Lucas the foundation of his anti-mechanistic argument: "Gödel's theorem must apply to cybernetical machines, because it is of the essence of being a machine, that it should be a concrete instantiation of a formal system." (Lucas (1961), p. 113).

Both historical connexions are important and strongly suggest a comparison of Galvan's conclusions – and especially of his concepts of "vision", intuition", and "understanding" of a Gödel's sentence being true "in the system" – with Searle's Chinese Room thought experiment, based on the concept of intentionality and highly critical of the computational theory of the mind. In my opinion, there is an element of truth in their claim, which I'm going to take up and develop, in a somewhat different way, and which, paradoxically enough, will make it consistent with the opposite element of truth contained in Turing's functionalism.

3 Turing, Searle, and the Chinese Room Thought Experiment

One the most important targets of Searle's Chinese Room thought experiment[5] was the Turing test, charged with "behaviourism" (Searle (1980a), p. 423) and, just

5 Cf. Searle (1980a), pp. 417–418. For the most concise formulations of this thought experiment that I have found, see Searle (1999), p. 115.

like strong AI in general, "functionalism". As far as behaviourism is concerned, Searle later said:

> [T]he Turing Test [...] is a straight expression of behaviorism. If it walks like a duck and talks like a duck, etc., then it is a duck, and if it behaves exactly as if it understood Chinese then it does understand Chinese. [...] If you accept the behavioristic criterion for the presence of the mental, then the mental is unlikely to be anything truly substantive of a biological nature. It is unlikely to be like digestion or photosynthesis or the secretion of bile, or any other natural human biological process. So the behaviourism of the Turing Test goes well with the idea that the mind is something formal and abstract. (Searle (2002), pp. 66–67)

According to Searle, this is intimately connected with the functionalist claim, which he also considers at the heart of the Turing test, that the system that implements the program is irrelevant and any hardware implementation "will do, provided only that it is rich enough and stable enough to carry the program." However, although the program enables the person of the Chinese Room thought experiment to pass the Turing test for understanding Chinese, the person in question "does not understand a word of Chinese" (Searle (2002), p. 51).

As Searle says, the main point of his thought experiment is that the purely formal, abstract or syntactical processes of an implemented computer program "could not by themselves be sufficient to *guarantee* the presence of mental content or semantic content of the sort that is essential to human cognition." (Searle (2002), pp. 51–59) He put his thought experiment also in the form of a logical argument: "1. Programs are formal (syntactical). 2. Minds have contents (semantic contents). 3. Syntax is not identical with nor sufficient by itself for semantics. From these we can derive: Programs are neither sufficient for nor identical with minds; i.e. strong AI is false." (Searle (1991), p. 526). For this reason, he points out that it is intrinsically impossible to duplicate intentionality *by computational means*.

In this connection may be mentioned the "Robot Reply". According to this objection, it is possible not only to develop programs consisting only in a set of rules that establish the permissible and necessary transitions from one state of the machine to another, but also programs with semantic import, which interact with the robot's environment.[6]

Against this Searle insisted over and over again that, if intentionality is defined as "the feature of certain mental states by which they are directed at or

6 Cf. Searle (1991), p. 526; cf. also Searle (1980a), p. 420. Among those who have endorsed versions of this objection, see especially Fodor (1980), Boden (1988), pp. 238–251, Crane (2003), p. 127–128, Melnyk (1996), p. 400, and Haugeland (2002), p. 386.

about objects and states of affairs in the world", the addition of "perceptual" and "motor" capacities does not add intentionality to any computer program (Searle (1980a), p. 424, footnote 3). To illustrate this point, he devises a variant of his thought experiment. Suppose that, instead of the computer, you put a homunculus inside the robot. All he does is manipulate formal symbols; he receives "information" from the robot's "perceptual" apparatus and gives out "instructions" to its motor apparatus without being aware of either of these facts. Unlike the traditional homunculus, he doesn't know what's going on (cf. Searle (1980a), p. 420).

From this quick sketch of Searle's view, it should be clear that, in so far as the philosophical consequences of Gödel's incompleteness theorems are concerned, what is common to both Searle's and Galvan's view is not only the intention to point out the shortcomings of the computational point of view, but also, more especially, to claim that understanding, semantic meaning, intuition, and in a word intentionality, are properties that man has but the machine lacks. However, there is a difference. In comparison with Galvan, who does not say anything about this point, Searle does not exclude the possibility in general of *duplicating* human intelligence. For him, there is no reason in principle why we could not build a machine able to perceive, act, understand, learn, etc., since our bodies with our brains are precisely such "intentional" machines; and it follows therefore that a machine can have intentionality if it does duplicate – either mechanically or in any other way – the biological structure of animal brains:

> We know that thinking is caused by neurobiological processes in the brain, but there is no logical obstacle to building a machine that could duplicate the specific causal powers of the brain to produce thought processes. The point, however, is that any such machine would have to be able to duplicate the specific causal powers of the brain to produce the biological process of thinking. The mere shuffling of formal symbols is not sufficient to guarantee these causal powers, as the Chinese room shows. (Searle (1999), p. 116; in the same sense, see also Searle (1980a), p. 417)

All the important objections to Searle's thought experiment were practically published along with Searle's original 1980a presentation, and now they exist in many different versions. Here I shall confine myself to pointing out Searle's wavering treatment of the nature of intentionality.

In different ways, and to different degrees, many authors have noticed the presence of a certain tension in Searle's thought (see, for example, Dennett (1987), p. 336; Jacquette (1989), pp. 618–619; Preston (1995), pp. 139–141; Melnyk (1996),

Corcoran (2001), Winograd (2002), p. 87). Moreover, that here is a general tension at the bottom of his thought we find explicitly stated by Searle himself:

> There is [...] an interesting tension. It is not at all easy to reconcile the basic facts with a certain conception we have of ourselves. Now, the question is, how we square this self-conception of ourselves as mindful, meaning-creating, free, rational, etc. agents, with a universe that consists entirely of mindless, meaningless, unfree, nonrational brute physical particles? In the end, perhaps we have to give up on certain features of our self-conception, such as free will. I see this family of questions as setting the agenda not only for my own work, but for the subject of philosophy for the foreseeable future. (Searle, 2006, p. 102)

On the one hand, intentionality is for Searle a "biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis or any other biological phenomena." (Searle (1980a), p. 424) Accordingly, Searle claims that intentionality, understanding, and meaning as such can be technically realized in a machine, provided it is made of biological stuff or imitates the specific biochemistry of the brain. On the other hand, Searle maintains that "the mind consists of qualia [...] right down to the ground" (Searle (1992), p. 20), that consciousness and intentionality can be understood only from the "the point of view of the agent, from my point of view" (Searle (1980a), p. 420) or from "the first person point of view" (Searle (1980b), p. 451; Searle (1992), p. 16), and that "the ontology of the mind is a first-person ontology" (Searle (1992), p. 20). This is a strong antinaturalistic tendency in Searle's thought and, notwithstanding Searle's complaints (cf. for example Searle (1992), p. xii), it seems to be inconsistent with the claim that intentionality is a "biological phenomenon", or, more in general, with the "biological naturalism" (Searle (1992), p. 1) he claims for his view.

Thus, that there is a tension in Searle's concept of intentionality there can be no doubt. Moreover, it is apparent that this wavering repeats and reflects the tension that according to Galvan exists between the two fundamental interpretations of Gödel's incompleteness theorems. Here, in a very generalized sense, we have in another form the same problem. But the changed conditions of the problem make its solution more significant for our purposes. What is the source of the mentioned tension in Searle's concept of intentionality? To answer this question throws light upon Searle's fundamental epistemological mistake and, at least in a certain sense, suggests a way out of Gödel's dichotomy, so far as this is related to Turing's way of putting the question whether machines can think.

On the one hand, Searle is at least in part aware of the *perspectival* character of intentionality or, more generally, of human knowledge. According to him, every intentional state has an aspectual shape, in the sense that it is directed at an object only "under an aspect". This applies just as much to conscious as to unconscious mental phenomena: "unconscious mental phenomena [...] to the

extent that they are genuinely *intentional* [...] must in some sense preserve their aspectual shape even when unconscious, but the only sense we can give to the notion that they preserve their aspectual shape when unconscious is that they are possible contents of consciousness." (Searle (1992), pp. 159–160).

On the other hand, as we have pointed out, Searle insists that intentionality is produced by the brain[7], and this claim, on reflection, is inconsistent with the thesis of the aspectual or perspectival character of intentionality because it sees the brain or some of its properties as an ultimate constituent of matter or reality. If an ultimate biochemical phenomenon existed independently of any particular point of view, if it existed in itself, apart from our knowledge interests, it would be a kind of atomic, self-enclosed reality. This sort of separateness does not and cannot exist. What we call the brain and its properties (including intentionality as a biological phenomenon) cannot exist or be understood apart from theoretical construals of some type. They appear as biochemical realities only through the concepts, terms and technical apparatuses that define the particular viewpoint from which biochemistry investigates reality. However, the perspective of biochemistry does not have any ontological priority over either that of physics or that of any other discipline.

Dennet remarked that Searle adopts, at least in part, the conceptual apparatus of Cartesianism (cf. Dennett (1987), p. 336, Hauser (2002), p. 129), but this is true in a different way from how he would expect. On reflection, to assume at the outset that intentionality is an empirical property resulting from causal-biological processes is only another way (not idealistic, but naturalistic) of hypostatising thought. And this is as gratuitous as, for example, the assumption of Driesch's entelechy. In both cases, the activity of consciousness is turned into a fact consisting of biological stuff. However, unlike the human brains, which one may only investigate by the usual methods of observation and experimentation, the devices of the laboratory will not help us to understand both Driesch's entelechy and intentionality in Searle's sense. Intentionality as a biological phenomenon is an empirical hypothesis about the causal origin of a fundamental power of the human mind. This hypothesis, however, is inferred from that which it presumes to explain, namely the possibility of ascribing meaning to real entities such as ink marks on paper. Obviously enough, if one admits that only organic (human, animal, and perhaps vegetal) stuff has intentionality, in the sense of purporting to represent objects, the falsehood of the Turing test may be regarded as tautological. But then the objection that this is question begging would be fatal: this is exactly the question to be answered.

7 Moreover, according to Searle, we have a biologically evolved innate capacity both for "individual" (cf. Searle (1983), p. 8) and for "collective intentionality" (Searle (1995), p. 37).

4 Two senses of intentionality and the truth of functionalism

As I have stressed, the perspective of biochemistry does not have any ontological priority over either that of physics or that of any other discipline. However, what is the point of view from which this judgment is made? Each of the special sciences has methods and concepts that reflect its point of view, but none of them has the means to answer the questions about the nature and conditions of its own kind of knowledge. What the concept "physics" means, is a question that does not belong to any special science, and especially not to physics, because the devices of the laboratory will not help us to answer this question. This business, consisting in the task of defining the special form of rationality of each science, belongs to philosophy of science, in which philosophy and science, though distinct in principle, cooperate and are really supplementary and inseparable.

From this point of view, it may be said that the mentioned tension in Searle's concept of intentionality arises from the opposition of science and philosophy: even though a particular, biological meaning of intentionality is regarded as fully compatible with a truly philosophical one, still they are placed by Searle, without mediation, side by side, as independent things that cannot cooperate with one another. If the essential opposition or antagonism of philosophy and science were accepted, nothing better than something similar to "Gödel's dichotomy" could result from our efforts: either the human mind has resources that are not at the disposal of the machine, and in particular the capacity to grasp contents that are abstract and not finite, or it can only gain insight through forms of finitistic evidence and there is no difference in principle between the human mind and the machine. However, as I shall tray to show, both alternatives are viable, though in different senses.

More precisely, the distinction between two senses of the term "intentionality", as a scientific (for example, biological) phenomenon on the one hand, and as a general philosophical concept, intimately connected with understanding and meaning, on the other, is a necessary but not sufficient condition both for escaping the tension in Searle's corresponding concept and, at least in a certain sense, for overcoming "Gödel's dichotomy".

What we need is something which will enable us to grasp the two senses of intentionality in their necessary relationship of distinction *and* unity.[8] A brief consideration of the *perspectival* character of intentionality will be sufficient

8 What I can say here is necessarily sketchy. For some more details, see Buzzoni (2013).

to directly support this connection, made up of both unity and distinction. On the one hand, as we have pointed out, aspectuality is an essential ingredient of the concept of intentionality: every intentional state has an aspectual shape in the sense that it is directed at an object only "under an aspect". "Aspects" or "aspectual shapes" here refer, roughly, to what spoke of in connection with seeing-as phenomena: e.g., seeing things as alike and seeing ambiguous pictures (e.g., Necker cubes, duck-rabbits) as one thing or the other. Seeing the duck-rabbit picture (cf. Wittgenstein (1958), II, xi) as a duck would (in these terms) be seeing its duck aspect or seeing it "under" its duck aspect.[9]

On the other hand, the definition of a particular (for example, biochemical) perspective presupposes its distinction from other particular perspectives such as those of physics, sociology, or of any other particular scientific discipline, and this distinction cannot be effected by means of the limited conceptions of any special perspective, but only from a wider and more comprehensive point of view. In other words, intentionality, as aspectuality, implicitly presupposes a different sense of intentionality, which is the very condition for understanding any determinate intentional act directed at any particular aspect of reality. More in general, we cannot be aware of the aspectual or perspectival character of intentionality (i.e. of the fact that all intentional acts depend upon certain adopted points of view) unless we are aware of our capacity in general to distinguish, change, abandon, invent, or to put in question old and/or novel points of view from which reality can be seen. Seeing a picture as a duck is seeing it under an aspect because we imagine, implicitly or explicitly, that it can have other different aspects, that it can be seen from other different angles (for instance, as a rabbit or as some lines on the sheet). In other words, to know a particular point of view in its particularity means to assume the capacity in principle to go beyond any particular point of view, to modify or to replace it by a new one that does not yet exist. This capacity to know reality from a potentially infinite (not determinable a priori) number of perspectives or theoretical points of view, or conversely this capacity to understand that every real object cannot be

9 As far as intentionality is concerned, apart from Husserl's concept of 'Abschattungen' and Wittgenstein's analysis of seeing-as phenomena (cf. Wittgenstein (1958), II, xi), this point has been more recently made especially by Haldane (1989), Putnam (1992), and McGinn (2004). Haldane maintained that any psychological description of human persons must contain, at least in part, the concepts by means of which they understand the world. Such a description is unavoidably intensional. Thus, since no sentence that is entirely extensional implies an intensional one, "there can be no deductive explanation of the emergence of intentional states" (Haldane (1989), p. 310). To think a always means to think a in one way or another and this determines the irreducible intensionality of intentional contexts (Haldane (1989), p. 312).

representatively exhausted even if we multiplied these points of view indefinitely, cannot itself be reduced to one particular point of view; in particular, it can be reduced neither to a computational (Turing) nor to a biochemical entity (Searle). Thus, the analysis of the oscillation in Searle's notion of intentionality leads us to distinguish, and at the same time to connect with one another, two senses of intentionality, namely a reflexive-transcendental (or pre-operational) and a positive point of view.

In order to cast a little more light on this matter, I shall connect it with the notion of thought experiment. As I have just said, the fact that every intentional state is directed at an object only "under an aspect" presupposes our capacity of modifying the available points of view or of inventing new ones from which to see reality differently. In other words, the awareness that the content of a particular intentional act would have been different if we had adopted a different point of view, in its turn, presupposes the capacity of the mind to assume that every *de facto* given reality could be different. The world appears to us as it does only because we can imagine, at the same time as we perceive it, that it might have been different. This radically counterfactual character of the mind, which can experience and acknowledge something as real only after having taken into consideration the possibility of its non-existence, is the wellspring of all hypotheses and of all particular thought experiments.

The ability of the mind to assume any reality hypothetically or counterfactually is, in Kant's terms, a transcendental fact of reason. It cannot be reduced to any particular intentional act relative to any particular object because it is the condition of the possibility of conceiving all particular hypotheses and all particular objects. In my opinion, what Searle calls "intentionality" or "semantic understanding", and what in Galvan appears as "understanding", "vision" or "intuition", is really nothing but another way of expressing the counterfactual nature of human reason.

A very important conclusion follows from this fundamental thesis when we try to answer the following question: Assuming that we have succeeded in building thinking beings or machines, which are endowed with the ability to assume any reality hypothetically or counterfactually, *how might we come to know that*? It is clear that, in order to find a solution to the problem at hand, Searle's thesis that the essence of thought and intentionality is chemical and biological has no advantage over the computational point of view; and perhaps it may even be regarded as a serious obstacle to finding an answer. In fact, to deduce that knowledge from the fact that these beings or machines behave as if they possessed intentionality in Searle's sense, is a genuine option for Turing but not for Searle. For we know that for Searle, behavioural indistinguishability is insufficient to establish intentionality: as we have seen, Turing's "mistake" lies in thinking that,

if "it walks like a duck and talks like a duck, etc., then it is a duck, and if it behaves exactly as if it understood Chinese then it does understand Chinese." (Searle (2002), p. 66) If this is accepted, to know that we have duplicated human intelligence would be a much more difficult undertaking than in the case of the Turing test if we had done just that by means of "an artefact" or "a man-made machine".

But the truth is that *we have no idea how a real material system, which can always occur and develop in only one way, could instantiate the hypothetical-reflexive domain of the mind, which can always contradict itself*. If we mean by "material reality" concrete, particular reality, which we can modify by means of our body either directly or through the mediation of instruments, we cannot understand how intentionality in its transcendental sense could be implemented in a machine. The possibility of robots made of biochemical stuff does not alter the substantial fact: biochemical stuff too always occurs and develops in only one way; it can neither contradict itself nor be counterfactual with respect to any other given reality.

This does not mean that from this point of view we cannot find an important element of truth as well in Searle's Chinese Room argument and in Galvan's notions of a "vision", "evidence" or "intuition", as in Turing's implicit functionalism. On the one hand, given the counterfactual nature of the mind, as we are acquainted with it in the first person, the only way to conjecture that there is the same 'intentionality' in ourselves as well as in other people is to start from empirical knowledge obtained by means of our body's interaction with the surrounding empirical reality. It is only when we succeed, at least to some extent and in certain respects, in reconstructing causal interactions in the first person that we come to guess intentionality in its transcendental sense and to pick out certain entities to which we attribute a mind, without falling into coarse or truly primitive anthropomorphism.

Obviously, not all intentional acts that we experience in ourselves are equally relevant to this process of ascribing intentionality to other people. Some intentional acts (for example, reading and writing, giving answers to questions (Turing's imitation game!), having new ideas, laughing over a joke, etc. seem to be more probably connected with the capacity of intentionality in its transcendental sense, that is to say with the capacity of making anything which is immediately given into a possibility or a problem. It is no accident that these kinds of behaviours are at the bottom of most 'Granny Objections' to the Turing test (cf. http://users.ecs.soton.ac.uk/harnad/Hypermail/Explaining.Mind96/0069.html).

On the other hand, strictly speaking, the intentionality in its pre-operational, pre-predicative or transcendental sense as well as the understanding of the truth of $G(T)$ in a formal system cannot be realized in any particular or concrete instantiation of whatever kind. Intentionality, in its pre-operational, pre-predicative

or transcendental sense both eludes all operational procedures that aim to grasp it entirely because is the condition of the possibility of all our particular intentional acts. Each concrete, particular ability may be added or subtracted in the Chinese Room thought experiment without altering the substantial fact concerning intentionality in its transcendental sense: *if* a robot did not already possess intentionality in this sense, there would be no way for him to gain it by accumulating particular abilities; *if* he did already possess it, he would continue to have it – at least in the form of pure possibility – whatever particular abilities he may lose. Even if a machine passed the Turing test in one of its most radical forms[10], this would still be insufficient to claim that intentionality and consciousness, in their transcendental sense, are encapsulated by the correct performance of computations.

That is another way of making the objection that Searle discussed under the heading of "The Other Minds Reply". In fact, the problem raised by the Turing test is intimately connected with the problem of other minds. To be more precise, it is essentially the same problem; indeed, when we are with other people, we are in principle in the same situation in which we would be if we were with machines that are able to successfully pass one of the generalized versions of the Turing test.

Searle's reply to the objection of "The Other Minds Reply" is very short: we simply assume that other persons have minds, just as in physics we assume the existence of the objects we deal with (Searle (1980a), p. 422). Admittedly, in most daily life contexts, the doubt concerning the existence of other people (whom we may regard, at least in a sense and hypothetically, as not subject to the Gödelian in principle limitation) may seem, and rightly so, abstruse and hard to understand. But first, even if this is admitted, the question remains how we come to have the concept of a mind that is not subject to Gödel's incompleteness theorems. Empirical-scientific considerations are necessary for this concept, but they are not sufficient, since they are relevant only if connected with the philosophical presupposition of intentionality in its transcendental sense. From particular intentional states we have to trace our way to the intentional-personal source which accompanies them as the condition of their possibility. We could not conceive of this concept if we did not presuppose in ourselves – in our thinking, acting,

10 See for example Erion (2001), p. 37; Harnad (1991), p. 44; Bringsjord et al. (2001). Some scholars have maintained that, even in its original form, the Turing test is already "too hard" and there is no need for even stronger versions of the Test (French (1990); on this point, see also French (1995) and French (2000)). In my opinion, the generalised versions go against the letter, but not the spirit of Turing's question and answer method, which was chosen so that the test could be in principle extended to any human activity that is relevant to a comparison between humans and machines (cf. Turing (1950), p. 433).

loving, etc. – an instantiation of the same universal concept of intentionality (understanding, intuition, vision, etc.). In other words, the belief that there are other minds presupposes that we already have the reflexive-transcendental concept of the mind, that we understand its counterfactual nature, its capacity of contradicting itself and its being the source of whatever particular intentional acts or operations we are able to carry out. Moreover, in all more or less doubtful cases, intentionality, with its self-transcendence and presence in absence, turns out to be a very evasive and mysterious process in everyday life too. In these cases, we are confronted with the question whether particular, apparently intentional abilities or performances warrant our treating their carriers as human persons.

Summing up, we may say that, contrary to what Searle believes, intentionality, taken in its reflexive-trascendental sense, is simply invisible to the scientific eye. A man and a machine or a robot that is and one that is not endowed with intentionality are no different for empirical science, since there is no empirically detectable difference between them that yields a definite criterion for intentionality in its transcendental sense.

This must be reconciled both with Turing's defence of machine 'intelligence' and Searle's (or Galvan's) argument to the opposite effect. In fact, the main strength of functionalism lies here. The Turing test is, so to speak, the other side of Wittgenstein's private language argument: "what is left over – Wittgenstein asked – if I subtract the fact that my arm rises from the fact that I raise my arm?" (Wittgenstein 1956, § 621) According to our previous account of intentionality, the answer must be: absolutely nothing *empirically detectable*. Exactly as Turing wanted to show, and as occurs in Wittgenstein's example of the beetle, "one can 'divide through' by the thing in the box; it cancels out, whatever it is."[11] And this, *pace* Searle, applies equally to a robot made with screws and bolts as to an android composed of organic matter. The question of the nature of the material basis necessary to carry intelligence and intentionality is here irrelevant. As far as this point is concerned, functionalism, interpreted in one of its more general senses according to which beings with different physiology or hardware could have the same types of mental states as humans, is in principle a tenable view, even in case empirical research took *de facto* very different paths. In principle, one

11 Wittgenstein (1956), par. 293. In this sense, Obermeier, who revisited Searle's thought experiment from a Wittgensteinian point of view, is right in saying "a Wittgensteinian account of 'understanding' does indeed support claims made by advocates of 'strong AI' that pertain to the role and function of language in the understanding process." (Obermeier (1983), p. 340). But this, for reasons already given, does not exclude intentionality in the transcendental sense.

can set no absolute limit to scientific research and its technological realizations, including the materials used therein.

A little reflection shows that this point of view is close to that of the "new mysterians", who assert the inexplicability of conscience and thought (cf. Flanagan (1991), p. 313. Cf. also the very similar claims made by authors as different as Fodor and McGinn (Fodor (1998), p. 83; McGinn (1991), p. 61). And it is no accident that, at least in a certain sense, even Turing could be considered a representative of this point of view. Turing wrote:

> I do not wish to give the impression that I think there is no mystery about consciousness. There is, for instance, something of a paradox connected with any attempt to localise it. But I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper. (Turing (1950), p. 447)

Turing's answer to what he called the "mathematical objection" should be interpreted in the light of this passage. Turing's answer certainly pointed out that what cannot be proved in a system can be proved in another formal system, but it also alludes to the fact that one must leave open the possibility that the limits of today's computers, whose architecture we know very well, may yet be transcended by human artefacts, the nature of which we cannot tell in advance, and the possibility of which – exactly as in the case of our mind – we cannot even comprehend.

This is not to deny that there is a fundamental divergence from Wittgenstein, Turing, and the functionalism of the early Putnam (and this same divergence indirectly provides an important point of contact with Galvan's conclusions). The inference from our particular abilities or operations to the ascription of intentionality in its reflexive sense is not a scientific inference, since it does not begin and end with empirical propositions as must be the case, at least in principle, in the experimental sciences. However, while scientific inquiry must simply ignore it because intentionality (consciousness, understanding, intuition, vision, etc.) in its transcendental sense falls outside the field of observation and experience, the philosophical reflection is able to signal but not to solve the "mystery" of intentionality (consciousness, understanding, intuition, vision, etc.). In a word, if the reflexive-transcendental sense of intentionality is constitutive of the nature of the mind, it is impossible to produce both scientific and philosophical "proofs" of the existence of other minds, or, what is the same, neither science nor philosophy can provide a generally reliable criterion for answering Turing's philosophical question in particular empirical circumstances (to borrow a metaphor from Heraclitus, the human person is similar to the oracle at Delphi, which neither utters nor hides his meaning, but shows it by signs). It is important to notice

that, although this conclusion sets a limit to the empirical sciences, in another sense it allows science to follow its own path, legitimately ignoring, as far as experience is concerned, any philosophical-transcendental prohibition against investigating and producing (historical-empirical) intentionality, intelligence or consciousness.

This brings me to a final point, to which, in a certain sense, all that has been said so far may be regarded as leading up. If, on the one hand, the scientific eye is blind to reflexive-transcendental intentionality, and, on the other hand, philosophy cannot definitely prove the *existence* of other minds, an important consequence follows: we need a practical principle, capable of finding a bridge across the gulf separating intentionality given in the first person and intentionality given as a power of the outer world.

The best candidate for a principle that satisfies this condition is perhaps the so called precautionary principle.[12] This principle, extrapolated from bioethical and ecological contexts, provides an important rule for deciding in doubtful cases, in particular regarding, so to speak, 'intermediate' beings, which could with certainty be placed neither within nor without the properly human realm. In these cases we need such a principle because are faced with the problem whether these beings should be treated, or fully treated, as human persons.

Generally, our decision as to how strong the evidence should be for accepting a hypothesis as validated depends upon the gravity, in an ethical sense, of the consequences that might depend upon our erroneously accepting or not accepting the hypothesis (cf. Rudner (1954)). The evasive and positively (i.e. empirically and/or logically) inaccessible nature of intentionality in its reflexive-transcendental sense, which characterizes the human mind, encourages the highest prudence. With regard to the present case, the precautionary principle urges us to include among human persons the greatest possible number of doubtful cases. This is surely one of the most important reasons why some people are treated as persons in spite of the fact that they would be unable to pass the Turing test even in some of their more simple or restricted versions.

In the same sense, the precautionary principle gives an important contribution to the question of whether, for instance by Turing's imitation game (and *a fortiori* by one of its generalizations), we would be morally obliged to attribute minds to machines. In particular circumstances or under particular conditions, the precautionary principle might morally oblige us to attribute minds to machines. In most cases, the passing of the Turing test should be sufficient

12 On the precautionary principle, cf. for example Gollier et al. (2000), Foster et al. (2000), and Fisher et al. (2006).

to orient our practical decisions and to provide a last bulwark against possible discrimination, at least if we apply the criterion underlying the Turing test, namely the reciprocal and symmetrical use of language, and we combine it with both the acknowledgment of the value of the human person and the precautionary principle. If some day we were able to build artificial life systems that took part in our conversations and cooperated with us on common projects, we would very probably be morally obliged to treat them as human persons. Such a possibility, no matter how improbable in the light of our current state of AI and neuroscience, cannot be excluded a priori, even though a final decision can be taken only in concrete circumstances.

And the same holds a fortiori for alien beings from another planet who might be able to cooperate with us; *differences in colour, genetic inheritance, in the fact of being made of flesh and blood or screws and bolts could hardly legitimize different decisions as to whether or not to treat these beings as human persons.* Here lies the moral truth and the emotional appeal of functionalism: we cannot help but feeling that, in Steven Spielberg's "A.I.", David, a "mecha" or mechanical boy of the future, even though not endowed with a human body, has thoughts and feelings like us: machines that pass the Turing test in such a strong version should be included among intentional entities.

Massimilano Carrara

Naïve Proof and Curry's Paradox

1 Introduction

In classical first order logic (FOL), *trivialism*, the truth of all sentences, and *explosion*, the derivability of any sentence, are obtained using the rule *ex contradictione quodlibet* (ECQ): $A, \neg A \vdash B$. The classical justification for ECQ rests on the alleged evidence that no contradiction can be true, evidence rejected in paraconsistent theories, in particular by dialetheists, who hold that there are dialetheiae, i.e. propositions that are both true and false.[1] Indeed, dialetheism maintains the thesis that there are true contradictions, i.e. true sentences of form $(A \wedge \neg A)$, called *dialetheiae*. More generally, they call *dialetheia* any sentence that is both true and false. In an extensive series of papers and books (see for example, Priest (1979), Priest (2001), Priest (2002), Priest (2006a), Priest (2006b)), Priest claims that the paradoxical sentences obtained from self-reference are *dialetheiae*.[2]

In standard natural deduction of FOL, ECQ can be derived using reductio ad absurdum (RAA) and other apparently non-problematic rules. It is a standard derived rule of FOL. But if you hold that there are dialetheiae, in order to avoid trivialism, RAA should be immediately rejected. Unfortunately, banishing RAA is insufficient to avoid trivialism: Curry's paradox, from which trivialism follows, can be generated without using RAA, but with just *modus ponens* (MP) and the derived rule of Absorption, i.e. ABS: $(A \rightarrow (A \rightarrow B)) \vdash (A \rightarrow B)$. In order to save dialetheism from trivialism, Priest adopts in the Logic of Paradox (LP) (1979) the material conditional, for which he rejects the general validity of MP.

The crucial problem is whether *trivialism* can follow even from logical principles that are dialetheistically correct. In this paper I concentrate, specifically, on a notion that Priest himself introduced in *The Logic of Paradox* (1979), i.e.

1 Priest uses the terms 'dialetheiae' and 'true contradictions' to indicate 'gluts'. propositions both true and false, a term coined by K. Fine in Fine (1975). For an introduction to dialetheism, see e.g. Berto (2007).

2 A discussion on the same topic is in Beall (2009), Colyvan (2009), and Weber (2010). For a short general introduction to the topic, see Murzi and Carrara (2015).

Massimilano Carrara: FISPPA Department – Section of Philosophy, Padua University

DOI: 10.1515/9783110529494-005

that of naïve proof, a notion amplified in his *Is Arithmetic Consistent?* (1994) and developed also in other texts (Priest (2006b)).

In *The Logic of Paradox*, Priest developed an argument, grounded in the notion of naïve proof, to the effect that Gödel's first incompleteness theorem would suggest the presence of dialetheiae in the standard model of arithmetic. Chihara (1984), Shapiro (2002), and others Berto (2009), for example, criticised the argument. Much of the criticism was directed against the notion of naïve proof itself, in particular against the thesis that everything that is naively provable is true. Surely, if the notion of naïve proof is understood as embracing all proofs performed by real working mathematicians, as Priest seems to suggest, the thesis is hardly tenable. The aim of the paper is to show that from a notion of *naïve proof* – dialetheically acceptable – *trivialism* follows.

Finally, I would like to point out here that part of Sergio Galvan's research was on naïve proof and Gödel's incompleteness theorems.[3] I hope this paper can help others recognise the importance of these arguments and their implications for logic and its philosophy. In so doing, I hope to follow some steps of Sergio Galvan's research.

2 Curry's paradox and its arithmetical formalisation

Curry's paradox belongs to the family of so-called paradoxes of self- reference (or paradoxes of circularity). In short, the paradox is derived from natural-language from sentences like (a):

(a) If sentence (a) is true, then Santa Claus exists.

Suppose that the antecedent of the conditional in (a) is true, i.e. that sentence (a) is true. Then, by MP, Santa Claus exists. In this way, the consequent of (a) is proved under the assumption of its antecedent. In other words, we have proved (a). Finally, by MP, Santa Claus exists.

Of course, we could substitute any arbitrary sentence for 'Santa Claus exists', for example, that 'you will win the lottery', etc., which means that every sentence can be proved: from Curry's paradox, *trivialism* follows. Priest (1979, IV. 5) observes that, in a semantically closed theory, using MP and ABS

ABS $(A \to (A \to B)) \vdash (A \to B)$

3 The notion of naïve proof and Gödel's incompleteness theorems have been developed by Sergio Galvan in particular in Galvan 1983, Galvan 1992, pp. 183–202).

a version of Curry's paradox is derivable. In what follows, I reconstruct his argument in the language of first order arithmetic with a truth predicate.[4] Let L be the language of first order arithmetic and N be its standard model. Extend L to the language L^* by introducing a new predicate T. With reference to a codification of the syntax of L^* by natural numbers, extend N to a model N^* of L^* by interpreting T as the truth predicate of L^*, so that, for all $n \in N$, $T(n)$ is true iff n is the code of a true sentence A of L^*, in symbols $n = \ulcorner A \urcorner$.

Of course, classically, such an interpretation is impossible, because the theory obtained by adding to Peano arithmetic (PA) the truth predicate for the extended language L^* with Tarski's biconditionals is inconsistent. Not so for dialetheism, where inconsistent models are accepted. But if one uses the classical rules of the conditional in natural deduction (from which ABS is derivable) and Tarski's scheme, i.e.:

$$T(\ulcorner A \urcorner) \leftrightarrow A$$

the model N^* turns out to be trivial. Let A be any sentence of L^*. By diagonalisation, there is a natural number k such that

$$k = \ulcorner T(\underline{k}) \rightarrow A \urcorner$$

We can derive A as follows:

1	(1)	$T(\underline{k}) \leftrightarrow \ulcorner T(\underline{k}) \rightarrow A \urcorner$	Tarski's schema
2	(2)	$T(\underline{k})$	Assumption
1, 2	(3)	$T(\underline{k}) \rightarrow A$	1, 2 MP
1, 2	(4)	A	2, 3 MP
1	(5)	$T(\underline{k}) \rightarrow A)$	2, 4 → Introduction
1	(6)	$T(\underline{k})$	1, 5 MP
1	(7)	At	5, 6 MP

Priest blocks this derivation in LP by rejecting the general validity of MP. According to him, this rule is not valid but *quasi-valid*, i.e. valid insofar as no dialetheia is involved. Priest (2006a), in order to reject MP, identifies, in the object language, $(A \rightarrow B)$ with $(\neg A \vee B)$. Then, the rejection proceeds as follows:

Proof. Suppose that A is a dialetheia; $(\neg A \vee B)$ is true even if B is not. In this case, if you infer B from A and $(A \rightarrow B)$, you get from true premises a not-true conclusion. This shows that MP may fail to preserve truth. Thus, the possibility of dialetheiae justifies the rejection of MP. qed

4 I follow here Carrara and Martino (2011), and Carrara et al. (2011).

In the next sections I first introduce a notion of *naïve proof* (§3), then I argue (§4) how it is possible to obtain Curry's paradox using it, without adopting MP.

3 On naïve proofs

Let us consider *naïve* proof, a notion introduced by Priest in his *The Logic of Paradox* (1979) in order to argue that Gödel's first incompleteness theorem would suggest the presence of dialetheiae in the standard model of arithmetic. Priest describes the *naïve* notion of *proof* as follows:

> Proof, as understood by mathematicians (not logicians), is that process of deductive argumentation by which I establish certain mathematical claims to be true. In other words, suppose I have a mathematical assertion, say a claim of number theory, whose truth or falsity I wish to establish. I look for a proof or a refutation, that is a proof of its negation. […] I will call the informal deductive arguments from basic statements *naïve proofs*. (Priest, 2006b, p. 40)

The alleged paradox should be suggested by the analogy of the familiar *informal proof* of Gödel's undecidable sentence G with the liar's paradox:

> As is clear to anyone who is familiar with Gödel's theorem, at its heart there lies a paradox. Informally the 'undecidable' sentence is the sentence 'this sentence is not provable'. Suppose that it is provable; then, since whatever is provable is true, it is not provable. Hence it is not provable. But we have just proved this. So it is provable after all (as well). (Priest (2006b), p. 237)

This argument is well known and widely discussed in the literature. Following Dummett, we can call it the *simple proof* argument (Dummett (1959)). It is worth noticing that the *simple proof* holds even dialetheically. Some people have maintained that this proof implicitly uses the consistency of PA, which, according to Gödel's second incompleteness theorem, is formally unprovable. Observe, however, *passim* that the proof at issue does not assume the consistency of PA, but by virtue of the fact that what is provable is true, dialetheically, consistency does not follow. Nor does the proof exploit RAA, but the tertium non datur, which dialetheically holds. The proof runs as follows:

Proof. G is provable or not provable. But if it is provable, then it is true and hence, as it says, (also) unprovable. In any case, it is unprovable and hence true. qed

Priest holds that the naïve notion of proof of an L-sentence is recursive so that the predicate 'P' of *naïve provability* is arithmetic. It follows that the relative Gödel's sentence G is a dialetheia. Priest's argument for the claim that the notion of

naïve proof is recursive rests on the observation, supported by Dummett, that any mathematical proof of a sentence is recognisable as such. So, the argument goes, given a sentence A and a finite sequence p of formulas, one can decide whether p is a proof of A or not. It follows, by Church's thesis, that the relation between a proof and its conclusion is recursive. Priest's conclusion presupposes the following:

(i) naïve proofs form a well-determined set codifiable by a set S of natural numbers, and
(ii) there is a mechanical procedure for deciding whether a number belongs to S or not.

I am not going to discuss the evidence for either (i) and (ii). I would like to just consider naïve proof as a dialetheically acceptable notion and see what happens in terms of self-reference paradoxes, specifically in the case of Curry's paradox.

4 Naïve proofs and Curry's paradox

The new version of Curry's paradox[5] here proposed is obtained without making use of MP. I just make use of the notion of *naïve poof*. Consider the extension L' of the language L of first order arithmetic, obtained by introducing a new predicate $P(x)$. Extend the standard model N of arithmetic to the model N', where P is interpreted as *naïve provability* for the language L' (with reference to a numerical codification of the syntax of L'). More precisely, $P(\ulcorner A \urcorner)$ means: It is naïvely provable that A is true in N.

In *Naïve Proofs*, Priest observes, "It is analytic that whatever is naïvely provable is true. Naïve proof is just that sort of mathematical argument that establishes something as true. And since this is analytic, it is itself naively provable […]" (Priest (2006b), p. 238). Moreover, he argues, "If something is naïvely proved then this fact itself constitutes a proof that A is provable" (Priest (2006b), p. 238).

On the basis of the above remarks, one can argue that:

(a) $P(\ulcorner A \urcorner) \rightarrow A$ is naively provable;
(b) If A is naïvely provable, then $P(\ulcorner A \urcorner)$ is naïvely provable.

5 A version of this paradox is in Carrara and Martino (2011) and Carrara et al. (2011).

Similarly, let us extend L to a language L^* by introducing a binary predicate $D(x, y)$. Then, we can extend the standard model N to the model N^* of L^*, where D is interpreted as the naïve deducibility relation for L^* (with reference to a codification of L^*). $D(x, y)$ means the following:

– y is naively deducible from x.

Or, in more explicit terms:

– There is a naïve proof that, assuming that x is true in N^*, leads to the conclusion that y is true in N^*.

Consider a natural *deduction* system. Rules of *elimination* and *introduction* for D, analogous to (a) and (b) stated above, are as follows:

(DE) From premises A and $D(\ulcorner A \urcorner, \ulcorner B \urcorner)$, one can derive B. The conclusion depends on all assumptions upon which the premises depend.

(DI) From premise B, depending on the unique assumption A, one can infer $D(\ulcorner A \urcorner, \ulcorner B \urcorner)$, discharging A.

Theorem. *From* (DE) *and* (DI), *trivialism follows.*

Proof. Let A be any L^*-sentence. By diagonalisation, we get a natural number k such that $k = \ulcorner D(\underline{k}, \ulcorner A \urcorner) \urcorner$. Using $\underline{k}$ as a name of $D(\underline{k}, \ulcorner A \urcorner)$, suppose that $\underline{k}$ is true. Since $\underline{k}$ says that A is deducible from $\underline{k}$ and deduction is sound, A is true. So we have proved A from the assumption $\underline{k}$. Hence $D(\underline{k}, \ulcorner A \urcorner)$, i.e. $\underline{k}$, is true. And, since deduction is sound, A is true. qed

A formal proof of A in natural deduction (where $\underline{k}$ is used again as a name of $D(\underline{k}, \ulcorner A \urcorner)$ is as follows.

1	(1)	$\underline{k}$	Assumption
1	(2)	$D(\underline{k}, \ulcorner A \urcorner)$	1, Identity
1	(3)	A	1,2 DE
	(4)	$D(\underline{k}, \ulcorner A \urcorner)$	1, 3 DI (discharging (1))
	(5)	$\underline{k}$	4, Identity
	(6)	A	4, 5 DE

Since A is arbitrary, N^* is trivial. But N^* differs from N only for the introduction of the relation of naïve deducibility: the arithmetical sentences of L are interpreted in N^* as in N. Therefore N is trivial as well.

5 Conclusion

In this paper, I took a notion of *naïve proof*, defended by Priest in his discussion of Gödel's theorem. I consider his characterisation of the notion *via* (a) and (b). By using it, a new version of Curry's paradox is proposed, obtained without making use of MP.

Acknowledgments: I would like to thank Enrico Martino because much of the theoretical work done in this paper was developed in cooperation with him, Ciro De Florio for inviting me to write a paper for this volume, for his patience, support and friendship, and, finally, an anonymous reviewer. I met Sergio Galvan when I was no longer a young boy (!). I was lucky to have had the opportunity to know him and to count on his friendship. It was a pity to have known him so late. I usually feel like one of David Foster Wallace's young fish, and I have the feeling that things would have been better for me if I had met Sergio earlier. Anyway, *this is water*.[6]

6 'There are these two young fish swimming along, and they happen to meet an older fish swimming the other way, who nods at them and says, "Morning, boys, how's the water?" And the two young fish swim on for a bit, and then eventually one of them looks over at the other and goes, "What the hell is water?"' (David Foster Wallace, *This is Water: Some Thoughts, Delivered on a Significant Occasion, about Living a Compassionate Life*. Little, Brown and Company, 2009, 1).

Roberto Festa and Gustavo Cevolani

Exploring and extending the landscape of conjunctive approaches to verisimilitude

Abstract: Starting with Popper, philosophers and logicians have proposed different accounts of verisimilitude or truthlikeness. One way of classifying such accounts is to distinguish between "conjunctive" and "disjunctive" ones. In this paper, we focus on our own "basic feature" approach to verisimilitude, which naturally belongs to the conjunctive family. We start by surveying the landscape of conjunctive accounts; then, we introduce two new measures of verisimilitude and discuss their properties; finally, we conclude by hinting at some surprising relations between our conjunctive approach and a disjunctive account of verisimilitude widely discussed in the literature.

We thank Theo Kuipers and Luca Tambolo for reading a first draft of this paper and providing useful feedback. The first author gladly recalls how he was strongly inspired by Sergio's concept of an "enlarged reason", according to which the sophisticated tools of modern logic, traditionally applied in the philosophy of mathematics and of science, can be fruitfully employed also in other areas, including epistemology, ontology, and rational theology. In particular, the present authors are both inclined to believe that the logical notions of probability and verisimilitude will reveal very useful in these areas of philosophical research. In support of this, the two quotations opening this paper intend to suggest that the concern for verisimilitude, or truth approximation, is shared within a wide spectrum of philosophical tendencies, from Christian theology to Marxism. A fascinating historical account of the origins of the concept of verisimilitude and the related development of realist and fallibilist views of knowledge – from Carneades and St. Augustine to Cusanus and Peirce, from Engels and Lenin to Popper – is given by Niiniluoto (1987, ch. 5).

Roberto Festa: Department of Humanistic Studies, University of Trieste, Italy; e-mail: festa@units.it. Roberto Festa gratefully acknowledges financial support from PRIN grant (20122T3PTZ) *Models and Inferences in Science*.

Gustavo Cevolani: IMT School for Advanced Studies Lucca and Center for Logic, Language, and Cognition (Turin); e-mail: g.cevolani@gmail.com. Gustavo Cevolani gratefully acknowledges financial support from the Italian Ministry of Scientific Research within the FIRB project *Structures and dynamics of knowledge and cognition* (Turin unit: *D*11*J*12000470001) and from the University of Turin and the Compagnia San Paolo within the project *Assessing information models: exploring theories and applications of optimal information search* (S1315RIC14L1CEG01/D16D15000190005).

DOI: 10.1515/9783110529494-006

> Human conceptions [...] are relative, but these relative conceptions go to compound absolute truth. These relative conceptions, in their development, move towards absolute truth and approach nearer and nearer to it.
>
> V. I. U. Lenin, *Materialism and Empirio-Criticism*, 1908

> [S]cience and theology [...] share a common conviction that there is truth to be sought. Although in both kinds of enquiry this truth will never be grasped totally and exhaustively, it can be approximated to in an intellectually satisfying manner that deserves the adjective 'verisimilitudinous', even if it does not qualify to be described in an absolute sense as 'complete'.
>
> John Polkinghorne, *Quantum Physics and Theology*, 2007

1 Introduction

Explicating verisimilitude has proved a challenging task since Popper first introduced the notion in 1963. After Popper's definition was shown to be untenable (Miller (1974), Tichy (1974)), logicians and philosophers of science have put forward a number of competing explications of what does it mean for a theory or hypothesis h to be closer to the truth than another one (for surveys, see Niiniluoto (1998) and Oddie (2014)). As a result, the conceptual landscape of different accounts of verisimilitude is now quite crowded. In the attempt to put some order in this landscape, verisimilitude theorists have recently devised alternative classifications of existing accounts of this notion (Zwart (2001); Zwart and Franssen (2007); Schurz and Weingartner (2010); Schurz (2011); Oddie (2013, 2014)). In this paper, we aim at exploring and extending what Schurz (2011) calls the "conjunctive" approach to verisimilitude (as opposed to the "disjunctive" one).

We proceed as follows. In section 2, we briefly survey the post-Popperian research program on verisimilitude and draw a pocket map of the landscape of conjunctive accounts of verisimilitude. Then, in Section 3 we focus on the "basic feature" approach to verisimilitude, which has been developed in some recent papers by the present authors[1]. We present two new measures of verisimilitude

1 For early versions and motivations, see Festa (2007a,b,c, 2009, 2011, 2012) and Cevolani et al. (2011) for a more detailed exposition; for discussion of some applications see Cevolani (2011, 2013, 2014a,b, 2015, 2016), Cevolani and Calandra (2010), Cevolani and Crupi (2015), Cevolani and Festa (2009), Cevolani et al. (2010, 2011, 2012)), and Cevolani et al. (2013). Note that Theo Kuipers' explication of "descriptive verisimilitude" (Kuipers, 1982) anticipates some of the key ideas of the basic feature approach.

grounded on our basic feature approach, the second being a generalization of the first, which in turn is a generalization of the original measure presented in our previous contributions. In Section 4, we conclude by hinting at some surprising relations between our measures and other well-known verisimilitude measures.

2 Conjunctive approaches to verisimilitude

We start by introducing a small amount of notation and terminology in section 2.1. In section 2.2, we present Popper's original definition of verisimilitude and the post-Popperian research program arising from its failure. We then focus on so called conjunctive approaches to truthlikeness in section 2.3.

2.1 A propositional framework for verisimilitude

We assume that "the world" is described by a propositional language L_n with n atomic propositions $a_1, \ldots, a_n$.[2] Within L_n, one can express 2^{2^n} logically distinct propositions, including the tautological and the contradictory ones; as usual, these are denoted $\top$ and $\bot$, respectively. Given two propositions h and g, h is said to be logically stronger than g when h entails g but g does not entail h (in symbols: $h \models g$ but $g \not\models h$). Figure 1 displays the $2^{2^n} = 16$ propositions of L_2–with p and q as atoms. We shall make use of this toy language to illustrate some features of the definitions of verisimilitude discussed in the paper, and to compare them.

Among the factual, i.e., neither tautological nor contradictory, propositions of L_n, some play a special role and deserve special mention. A basic proposition is an atom or its negation (e.g., p, $\neg p$, q, $\neg q$ are the basic propositions of L_2 in Figure 1). The notation $\pm a_i$, where "$\pm$" can be "$\neg$" or nothing, will be employed to denote an arbitrary basic proposition of L_n.

A conjunction $\pm a_1 \wedge \ldots \wedge \pm a_m$ of m basic propositions ($0 \leq m \leq n$), at most one for each atomic one, will be called a *conjunctive proposition* of L_n. If $m = 0$, then the conjunctive proposition is tautological; if $m = 1$, it is a basic proposition; and if $m = n$, it is a so called *constituent* of L_n. Note that the $q = 2^n$ constituents $z_1, \ldots, z_q$ are

2 One may argue that an adequate explication of verisimilitude should not be restricted to theories stated in simple propositional languages. Still, as shown for instance in the works quoted in the previous note, verisimilitude measures for propositional theories prove both adequate and fruitful in the analysis of some relevant issues in formal epistemology and the philosophy of science

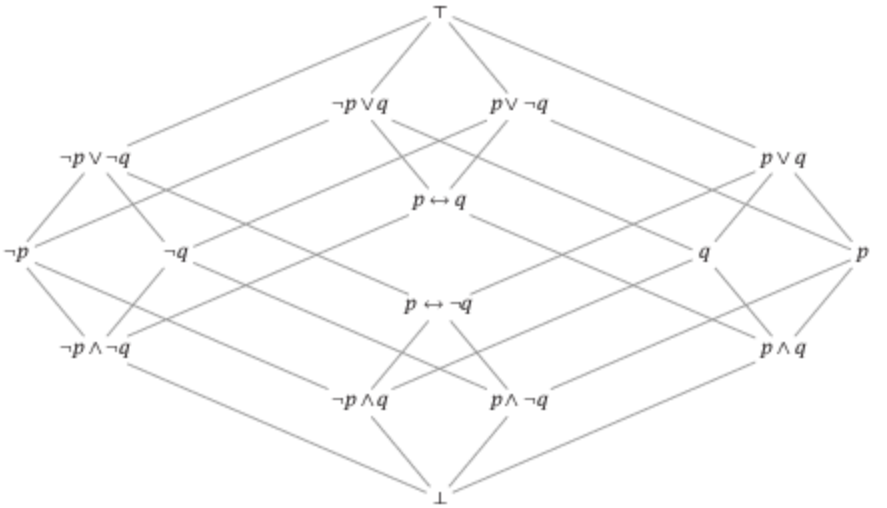


Fig. 1. The 16 (logically distinct) propositions of language L_2 (with atoms p and q) represented in increasing order of logical strength, from the top to the bottom of the diagram: if two propositions are (directly or indirectly) connected, the upper one is a consequence of the lower one.

the logically strongest factual propositions of L_n. As Figure 1 shows, constituents are weaker than a contradiction but stronger than any other proposition (the constituents of L_2 are the four conjunctions $p \wedge q$, $p \wedge \neg q$, $\neg p \wedge q$, and $\neg p \wedge \neg q$). The negation of a constituent, i.e., a disjunction of the form $\pm a_1 \vee \ldots \vee \pm a_n$, is called by Carnap (1950b, p. 405) a content element of L_n (in Figure 1, $p \vee q$, $p \vee \neg q$, $\neg p \vee q$, and $\neg p \vee \neg q$ are the four content elements of L_2). These are the weakest factual propositions of L_n, stronger than a tautology but weaker than any other proposition.

Note that each constituent is logically incompatible with any other, and that only one of them can be true; the true constituent is denoted t and is the strongest true proposition expressible in L_n. Intuitively, a constituent completely describes a possible world, i.e., a possible state of affairs of the relevant domain; thus, t can be construed as "the (whole) truth", i.e., as the complete true description of the actual world in $\mathcal{L}_n$. When one of the constituents of L_n is identified with the truth t, it partitions the set of propositions of L_n into the class $T = Cn(t)$ of the true ones and its complement F, containing the false ones. In the following, we shall assume, for the sake of illustration, that $p \wedge q$ is the truth of the toy language in Figure 1.

The *specular* of a conjunctive proposition $\pm a_1 \wedge \ldots \wedge \pm a_m$ is the conjunction of the negations of all its basic propositions $\pm a_i$.[3] As an example, in Figure 1 $\neg p \wedge \neg q$ is the specular of the truth $p \wedge q$. In general, the specular of the truth is denoted f. Intuitively, f can be construed as the "worst" constituent of L_n, i.e., as the completely false description of the actual world. Note the difference between the specular of the truth f – i.e., $\neg p \wedge \neg q$ in Figure 1 – and the negation of the truth $\neg t$, which is the only false content element of L_n – i.e., $\neg p \vee \neg q$.

Any proposition h of L_n is construed here as expressing a possible theory or hypothesis about the world. Intuitively, the verisimilitude of h depends on how much true information h provides about the world. In this connection, let $Cn(h) = \{g : h \vDash g\}$ be the class of propositions entailed by h (where Cn denotes the operation of classical logical consequence), i.e., what Popper (1963b, p. 218) called the "logical content" of h. For our purpose, it will be useful to consider also the "basic content" of h, i.e., the set $B(h) = \{\pm a_i : h \vDash \pm a_i\}$ of the basic propositions entailed by h or, as we may say, of the basic consequences of h. Of course, for any h, $B(h) \subset Cn(h)$, i.e., all basic consequences of h are consequences of h.

Finally, few words about probability. A probability measure m defined on the propositions of L_n is called a *logical* probability measure when it assigns to each constituent z_i of L_n the same value $m(z_i) = 1/2^n$ (cf. Carnap 1950b, ch. 5). For any proposition h of L_n, $m(h)$ is the proportion of constituents entailing h out of the total number of constituents:

$$m(h) = \sum_{z_i \vDash h} m(z_i) = \frac{|\{z_i : z_i \vDash h\}|}{2^n} \tag{1}$$

It follows that all basic propositions have the same degree of logical probability:

$$m(\pm a_i) = 1/2 \tag{2}$$

3 This notion of specularity was first introduced, as far as we know, by Festa in an unpublished 1982 manuscript; a summary of his results was then provided by Niiniluoto (1987, p. 319–321). Roughly in the same years, Oddie (1986, pp. 49–50) introduced the term "reversal" to denote the same concept (defined for arbitrary propositions, not just conjunctive ones). The term "inverse" has then been used later by Zwart (2001, p. 25) to refer to the same notion. In this connection, one should note that Zwart (*ibidem*, pp. 32 ff. and 56 ff.) discusses also a "specularity property" and acknowledges, following Kuipers (1987b, p. 85), that "the term is Roberto Festa's" (*ibidem*, note 66); however, he doesn't mention that, despite the common terminology, Festa's notion of "specular" and Kuipers' specularity property are actually quite unrelated (cf. Festa (1987, p. 153 ff)).

Assuming that h is consistent (as we shall always assume in the following), the conditional logical probability of g given h is defined as usual, i.e., $m(g|h) = m(h \wedge g)/m(h)$. This is the proportion of the cases (i.e., constituents) in which g is true out of the total number of cases in which h is true. If $m(g|h) > m(g)$, i.e., if h raises the initial proportion of cases in which g is true, it is customary to say that h is positively *relevant* for g.[4] Following Salmon (1969, p. 63), if h is positively relevant for g we shall say that h partially entails g or, equivalently, that g is a partial consequence of h. Note that, if h (fully) entails g, then g is true in all cases where h is true; in other words, any non-tautological consequence of h is also a partial consequence of h.

It also follows immediately that, as far as basic propositions are concerned, h partially entails $\pm a_i$ just in case $m(\pm a_i|h) > 1/2$. Accordingly, we call $b(h) = \{\pm a_i : m(\pm a_i|h) > 2\}$ the set of partial basic consequences of h. To illustrate, one can check, with reference to Figure 1, that the partial basic consequences of, say, $p \vee \neg q$ form the set $b(p \vee \neg q) = \{p, \neg q\}$. Of course, any basic consequence of h – i.e., any basic proposition "fully" entailed by h – is also a partial basic consequence of h: in other words, we have that $B(h) \subset b(h)$.

2.2 The post-Popperian research program on verisimilitude

In the early sixties of the past century, the controversy between Popper's and Carnap's followers concerning the goals of science and the growth of knowledge urged Popper (1963b, ch. 10) to introduce the first formal explication of verisimilitude. Popper believed that science aims neither at highly probable nor at inductively well confirmed theories, but at theories with a high degree of verisimilitude, a notion which "represents the idea of approaching comprehensive truth [and] thus combines truth and content" as the two fundamental cognitive goals of inquiry (Popper, 1963b, p. 237).

In Popper's intentions, the idea of verisimilitude should have supported his realist and falsificationist views about science, by showing how it is possible to keep together two apparently opposite tenets: i.e., that our best theories are bold conjectures which are likely false (and will be quite surely falsified in the future) and that still science progresses toward truth. If a false theory can be closer to the truth than another false theory, Popper argued, then one can coherently maintain

4 For a detailed analysis of the concepts of positive and negative relevance see Carnap (1950a, cg. 6) and Salmon (1969).

a falsificationist attitude in methodology and a realist view of the main aim of science, i.e., truth approximation.[5]

In order to defend the ideas outlined above, Popper (1963b) introduced a definition of verisimilitude based on an apparently very sound intuition: the more true propositions and the less false propositions a theory h entails, the greater its verisimilitude. More precisely, recalling that $Cn(h)$ is the class of consequences of h, let $Cn_T(h) = Cn(h) \cap T = Cn(h) \cap Cn(t)$ denote the class of true consequences of h, and $Cn_F(h) = Cn(h) \cap F$ the class of false consequences of h, to the effect that $Cn_T(h) \cup Cn_F(h) = Cn(h)$. Then, according to Popper (1963b, p. 233), h is closer to the truth than g if and only if h has no less true consequences than g (and possibly more) and no more false consequences (and possibly less).

Definition 1 (Popperian verisimilitude). *h is at least as close to the truth as g iff:*

$$Cn_T(h) \supseteq Cn_T(g) \text{ and } Cn_F(h) \subseteq Cn_F(g)$$

Moreover, h is closer to the truth than g if at least one of the above inclusion relations is strict.

When Definition 1 first appeared in the tenth chapter of *Conjectures and Refutations*, it didn't attract much attention, perhaps because most readers found the definition exactly as it should be (cf. Kuipers 2000, p. 139). Popper's definition became famous about ten years later, when Miller (1974) and Tichý (1974) independently proved that it was completely inadequate. More precisely, they showed that no false theory h can be closer to the truth than another (true or false) theory g according to Popper's Definition 1. This so called Tichý-Miller theorem proved fatal for Popper's explication of verisimilitude, since it showed that Definition 1 is worthless for the very purpose for which Popper introduced it – i.e., ordering false theories according to their closeness to the truth.

The surprising failure of Popper's attempt urged logicians and philosophers of science to develop more adequate definitions of verisimilitude. As a result, the conceptual landscape of different accounts of verisimilitude is now very crowded, different scholars having put forward a number of competing and partially conflicting explications of what does it mean for a theory to be closer to the truth than another one (for an early collection, see Kuipers (1987a) and, for surveys, see Niiniluoto (1998) and Oddie (2014)). At the moment, the work in

5 For most recent critical discussion about this view of scientific progress, see, e.g., Cevolani and Tambolo (2013a), Niiniluoto (2014) and Rowbottom (2015).

this area follows three different paths. First, the search for adequate explications of verisimilitude is still an active area of study, as the contributions by, e.g., Festa (2007a,b,c), Schurz and Weingartner (2010), Cevolani et al. (2011), Northcott (2013), and Kuipers (2015) testify. Second, such explications are being applied to the analysis of both classical problems in the philosophy of science (see, e.g., Cevolani and Tambolo (2013a,b), Cevolani et al. (2013), Tambolo (2015), Niiniluoto (2014, 2015) on the analysis of scientific progress) and of issues in formal epistemology or even cognitive psychology (see, e.g., Cevolani et al. (2010, 2012); Cevolani and Crupi (2015); Cevolani and Schurz (2017) on the analysis of paradoxes of rational belief, and Cevolani and Calandra (2010), Cevolani et al. (2011), Kuipers (2011), [Niiniluoto 2011], and Schurz (2011) on the connections between verisimilitude and belief revision). Finally, verisimilitude theorists have recently devised alternative classifications of existing accounts of this notion, in order to investigate the differences, similarities, and possible connections among the different approaches (Zwart (2001); Zwart and Franssen (2007); Schurz and Weingartner (2010); Schurz (2011); Oddie (2013, 2014)). Our point of departure is this recent debate on the most appropriate way to classify accounts of verisimilitude, and in particular the distinction between "conjunctive" and "disjunctive" approaches, to which we now turn.

2.3 Mapping the landscape of conjunctive accounts of verisimilitude

In some recent papers, Gerhard Schurz has convincingly argued that the way in which theories are represented in the first place can have significant implications in assessing their verisimilitude (Schurz and Weingartner (2010), Schurz (2011)). More precisely, he distinguishes two approaches to theory representation. Within the first, a theory h is represented as a conjunction of minimal "content parts", i.e., of the smallest items of information provided by h on the world; as an example, h may be the conjunction of its consequences in some language. The second approach instead represents h as a disjunction of maximal "alternative possibilities", like the possible worlds or the models of the underlying language. Of course, the above distinction parallels the familiar one between two equivalent ways of expressing sentences in formal languages, namely, the one between conjunctive and disjunctive normal forms.[6]

6 As immaterial as this distinction may be from a purely logical point of view, it can have significant implications for the formal analysis of epistemological concepts, including verisimilitude, as Schurz and Weingartner (2010, p. 424) observe (cf. also Carnap (1950a, §72–73), especially p. 407).

Within both approaches, one can distinguish different accounts of verisimilitude, according to the different ways of construing the relevant notion of either content part or alternative possibility. Here, we shall focus only on the conjunctive approach. Schurz (2011, csec. 2.2) surveys five conjunctive accounts of verisimilitude proposed in the literature, including our own basic feature approach, to be discussed in greater detail in Section 3. All these accounts retain the following fundamental Popperian intuition:

> *h* is at least as close to the truth as *g* iff
> all true content parts of *g* are also true content parts of *h* and
> all false content parts of *h* are also false content parts of *g*.

However, they differ on how these content parts are defined. The following list displays, in order of appearance in the literature, a number of conjunctive accounts, including the ones identified by Schurz (2011).

- In Popper's account (cf. Definition 1), the content parts of *h* are arbitrary logical consequences of *h*; as mentioned, such an account is untenable due to the Tichý-Miller theorem. All other accounts mentioned below eschew his problem.
- In Mortensen's (1978; 1983) account, classical logic is abandoned in favour of a relevant logic, to the effect that not all classical consequences of *h* count as relevant consequences of *h*.
- In the "short theorems" account (Mott, 1978), the content parts of *h* are special consequences of *h*, comparable to, but different from, the relevant consequences in Schurz's and Gemes' accounts below.
- In the "relevant element" account (Schurz and Weingartner, 1987, 2010; Schurz, 2011), the content parts of *h* are relevant consequences of *h*, according to the definition of relevance developed by Schurz in a number of papers (see especially Schurz 1991).
- In Gemes (2007)' account, the content parts of *h* are also relevant consequences of *h*, but the notion of relevance is different from the one employed by Schurz.
- In the "basic feature" account (e.g., Cevolani et al. 2011), the content parts of *h* are the "basic" consequences of *h*, i.e., the basic propositions entailed by *h*.
- In the "Carnapian" account (Cevolani 2016), the content parts of *h* are the content elements (in the sense of Carnap 1950b, p. 405) entailed by *h*, i.e., its weakest factual consequences.

With the exception of Popper's one, all the above accounts have a trait in common: the set of content parts of *h* is a proper subset of the class of its

logical consequences (cf. Schurz 2011, p. 206). This means that only some of the consequences of h are deemed relevant as far as verisimilitude assessments are concerned. In all cases, this is sufficient to avoid the unwelcome consequences of the Tichý-Miller theorem. Moreover, all conjunctive accounts (including Popper's one) meet what Oddie (2013, p. 1651) calls the "strong value of content for truths" principle, i.e, the requirement that, among truths, verisimilitude increases with content:

> if h and g are true and h is logically stronger than g,
> then h is more verisimilar than g.

This principle – or at least its weaker version, saying that if h is true and entails g, then h is at least as close to the truth as g – is regarded by most verisimilitude theorists, including Popper himself, as "an essential desideratum for any theory of verisimilitude" (Oddie, 2014, sec. 1). For these and other reasons, Schurz and Weingartner (2010, sec. 3) defend conjunctive accounts as intrinsically plausible and delivering cognitively more manageable assessments of verisimilitude.

Interestingly, the landscape of conjunctive approaches seems to be conceptually delimited by two extreme positions. The first is represented by Popper's original definition; the second is the newly introduced "Carnapian" definition of verisimilitude (see Cevolani 2016, for details). While for Popper verisimilitude depends on the set of all the consequences of h, assessments of Carnapian truthlikeness are based only on the set of the weakest consequences of h, i.e., the content elements entailed by h. Between these two extremes, one can arguably place all other conjunctive accounts, according to the different classes of consequences of h they isolate as relevant for verisimilitude comparison. Notably, both the extremes are inadequate as accounts of verisimilitude, but for different reasons: the Popperian account because of the Tichý-Miller theorem, and the Carnapian account since it meets the implausible condition according to which verisimilitude increases with logical strength (not only among true but also) among false theories:

> if h and g are false and h is logically stronger than g,
> then h is more verisimilar than g.

Such condition, that Oddie (2013, p. 1654) has dubbed "the strong value of content for falsehoods", is rejected by virtually all verisimilitude theorists, the only exception being David Miller, whose favoured account of verisimilitude satisfies it (cf. Miller 1978, 1994, 2006).

3 The basic feature approach to verisimilitude

In this section, we focus on our own conjunctive account of verisimilitude of propositional theories, i.e., the basic feature approach. We proceed in two steps, presenting in turns two variants of this approach. In the first, the verisimilitude of h depends on the *categorical* information that h provides about the basic features of the world (section 3.1). A discussion of the limitations of this version then leads to a second, refined one, according to which the verisimilitude of h is measured in terms of the partial information provided by h about the basic features of the world (section 3.2).

3.1 Verisimilitude as categorical information about the basic features of the world

The key intuition underlying most explications of verisimilitude can be expressed as follows: a theory h is verisimilar when h tells many things about the world and many of these things are true. In this sense, as Popper (1963b, p. 237) noted, the idea of verisimilitude "combines truth and content": h has to provide much information about the world, and most of this information has to be true, in order to make h (highly) verisimilar. Within the basic feature approach, the verisimilitude of h only depends on what h says about the basic features of the world. These are n independent facts which may or may not obtain in the world (like "it's raining" and "it's not raining") and are described by the atomic propositions of L_n. Accordingly, the key intuition above can be rephrased as follows: h is verisimilar when h provides much information about the basic features of the world and most of this information is true (cf. Cevolani et al. 2011).[7]

An immediate refinement of the above intuition concerns quantitative verisimilitude, i.e., the definition of an appropriate measure of the verisimilitude of a theory h. We will assume that such measure depends only on the *amount* of true and false information that h provides about the basic features of the world. More precisely, we require that the verisimilitude of h is an increasing function of the amount of true information and a decreasing function of the amount of false information provided by h on those basic features. There are many different ways

7 If rather sparse, consonant intuitions are recurrent in the literature on verisimilitude; cf., e.g., Brink and Heidema (1987, sec. 4); Oddie (1987, sec. 2), Kuipers (1982). Interestingly, similar ideas also appear in the field of "veristic social epistemology" (cf. Goldman (1999, sect. 3.4).

of specifying such a function; two of them will be discussed in this and the next section.

Recalling that $B(h) = \{\pm a_i : h \models \pm a_i\}$ is the set of the basic consequences of h, its cardinality $|B(h)|$ arguably provides a simple measure of the total amount of information provided by h about the n basic features of the world. In fact, this is equivalent to saying that (i) the amount of information provided by h about $\pm a_i$ is 1 if h entails either a_i or its negation $\neg a_i$, and is 0 in the case where h entails neither of them; and, (ii) the total amount of information provided by h about the n basic features of the world is just the sum of the amount of information provided by h about each of them. By dividing this number of basic consequences of h by n, the following normalized measure is obtained:

$$Inf(h) \equiv \frac{|B(h)|}{n} \tag{3}$$

As one can check, $Inf(h)$ varies between the minimum information provided by a tautology and the maximum information provided by a constituent:

$$Inf(\top) = 0 \leq Inf(h) \leq 1 = Inf(z_i) \tag{4}$$

Using now Popper's definitions (presented in Section 2.2) as a benchmark, we shall say that $B_T(h) = B(h) \cap T$ is the class of true basic consequences (or of basic truths) of h, and $B_F(h) = B(h) \cap F$ the class of its false basic consequences (or basic falsehoods). Accordingly, the amount $Inf_T(h)$ of true information provided by h about the n basic features of the world may be defined along the same lines of definition (3):

$$Inf_T(h) \equiv \frac{|B_T(h)|}{n} \tag{5}$$

In the same way, the amount $Inf_F(h)$ of false information provided by h about the n basic features of the world is defined as:

$$Inf_F(h) \equiv \frac{|B_F(h)|}{n} \tag{6}$$

It is easy to check that the information $Inf(h)$ provided by h is the sum of the true and false information provided by h:

$$Inf(h) = Inf_T(h) + Inf_F(h) \tag{7}$$

Interestingly, a simple measure of the verisimilitude of h is obtained from (7) by replacing the "plus" sign with the "minus" one:

$$Vs(h) \equiv Inf_T(h) - Inf_F(h) \tag{8}$$

In words, the verisimilitude of h is the difference between the amount of true and false information provided by h about the basic features of the world.[8]

The following inequalities are immediate consequences of definition (8):

$$Vs(f) = -1 \leq Vs(h) \leq 1 = Vs(t) \tag{9}$$

$$Vs(\top) = 0 \tag{10}$$

Note that (9) says that $Vs(h)$ varies between −1, i.e., the verisimilitude of the complete falsehood, and 1, i.e., the verisimilitude of the truth. Equality (10) shows that the verisimilitude of a tautology is a sort of natural middle point: $Vs(h) > 0$ iff the number of basic truths exceeds the number of basic falsehoods of h, while $Vs(h) < 0$ iff the number of basic falsehoods exceeds the number of basic truths of h.

The definition of verisimilitude just presented is essentially identical to the one proposed in our earlier work as limited to the class of conjunctive theories. For this kind of theories, as we argued in those earlier contributions, Vs provides perfectly adequate assessments of verisimilitude. Here, definition (8) is given for arbitrary theories, so that Vs is intended to measure the verisimilitude of both conjunctive and non-conjunctive ones. However, it is easy to show that, as a general measure of the verisimilitude for propositional theories, Vs is inadequate.

To see this, let us define the "conjunctive counterpart" of h, denoted $c(h)$, as the strongest conjunctive statement entailed by h or, equivalently, as the conjunction of the basic propositions entailed by h. It is now easy to check that, according to Vs, the verisimilitude of h is equal to the verisimilitude of its conjunctive counterpart:

$$Vs(h) = Vs(c(h))$$

This already raises a problem for Vs. In fact, it is clear that two theories h and g may have the same conjunctive counterpart even if they are not logically equivalent. As an example, one can check that any possible disjunction of basic propositions has the same, tautological counterpart:

$$c(\pm a_1 \vee \pm a_2) = c(\pm a_1 \vee \pm a_2 \vee \pm a_3) = c(\pm a_1 \vee \pm a_2 \vee \ldots \vee \pm a_n) = c(\top) = \top.$$

This fact is sufficient to show that Vs is a very coarse grained measure, since it assigns the same degree of verisimilitude to significantly different theories. In

8 As an immediate consequence of equalities (7) and (8), we find that $Vs(h)$ can be expressed in terms of $Inf(h)$ and $Inf_F(h)$ or, alternatively, in terms of $Inf(h)$ and $Inf_T(h)$: $Vs(h) = Inf(h) - 2Inf_F = 2Inf_T - Inf(h)$.

particular, if h is non-tautological but so weak that it doesn't entail any basic proposition, then h is assigned the same verisimilitude as the tautology, namely 0. For instance, whatever the truth t of L_n, one obtains that:

$$Vs(\pm a_1 \vee \pm a_2) = Vs(\pm a_1 \vee \pm a_2 \vee \pm a_3) = Vs(\pm a_1 \vee \pm a_2 \vee \ldots \vee \pm a_n) = Vs(\top) = 0.$$

As another example, with reference to Figure 1, all the content elements of L_2, as well as $p \leftrightarrow q$ and $p \leftrightarrow \neg q$, are deemed as truthlike as a tautology.

The equalities above show that *Vs* doesn't deliver intuitively sound verisimilitude assessments for non-conjunctive theories. Fortunately, as we argue in detail in section 3.2 below, there is a natural way of defining a verisimilitude measure for arbitrary theories which improves on *Vs* under this respect and is still based on the fundamental insight of the basic feature approach. To recall, according to this approach verisimilitude depends on how much true information h provides about the basic features of the world. Until now, we have worked with a "categorical" notion of information: h provides information about $\pm a_i$ just in case h logically entails $\pm a_i$; otherwise, h doesn't provide any information at all. Such notion, however, is too restrictive: intuitively, it is clear that, for instance, $p \vee q$ does provide at least some information about p, although not so much information as that provided by p itself. On the contrary, definition (3) implies that $p \vee q$ provides zero information about p, exactly as a tautology does: in short, a categorical account of information is too crude to deliver a fine grained definition of verisimilitude. To this purpose, we need a more graded, non-categorical notion of information, according to which the information provided by h on $\pm a_i$ is just the degree to which h entails $\pm a_i$. Such notion of "partial" information is introduced in the next subsection.

3.2 Verisimilitude as partial information about the basic features of the world

Recalling that $\pm a_i : m(\pm a_i|h) > \frac{1}{2}$ is the set of partial basic consequences of h, a simple definition of the amount of information provided by h about each of its partial basic consequences is the plain difference $m(\pm a_i|h) - \frac{1}{2}$. Intuitively, such difference measures the "distance" between the conditional logical probability $m(\pm a_i|h)$ and the absolute logical probability $m(\pm a_i)$. By multiplying the above expression by 2, one obtains the normalized measure

$$inf_i(h) = 2 \times (m(\pm a_i|h) - 1/2) \tag{11}$$

Note that $inf_i(h)$ is always positive and takes 1 as maximum value, when h entails $\pm a_i$.[9] By summing up the information provided by h about each of its partial consequences, one obtains the total amount of information provided by h on the basic features of the world, which can be normalized dividing by n:

$$inf(h) \equiv \frac{1}{n} \sum_{\pm a_i \in b(h)} inf_i(h) \tag{12}$$

Again, one can easily see that $inf(h)$ varies between the minimum information provided by a tautology and the maximum information provided by a constituent:

$$inf(\top) = 0 \leq inf(h) \leq 1 = inf(z_i) \tag{13}$$

Let us now denote $b_T(h) = b(h) \cap T$ the class of basic truths partially entailed by h, and $b_F(h) = b(h) \cap F$ the class of basic falsehoods partially entailed by h. Then, the amount $inf_T(h)$ of true partial information provided by h about the n basic features of the world is defined as:

$$inf_T(h) \equiv \frac{1}{n} \sum_{\pm a_i \in b_T(n)} inf_i(h) \tag{14}$$

i.e., as the normalized amount of information provided about the basic truths partially entailed by h. Similarly, the amount $inf_F(h)$ of partial false information provided by h about the n basic features of the world is defined as the normalized amount of information provided about the basic falsehoods partially entailed by h:

$$inf_F(h) \equiv \frac{1}{n} \sum_{\pm a_i \in b_F(n)} inf_i(h) \tag{15}$$

It should be clear that the above definitions are structurally identical to those given in the previous section for the categorical case (in particular, definitions (12), (14), and (15), on the one hand, are the counterparts of definitions (3), (5), and (6), on the other hand). For this reason, it is not surprising that similar considerations can be repeated here for the case of partial information. First, it is easy to see that $inf(h)$ is the sum of the true and false information provided by h:

$$inf(h) = inf_T(h) + inf_F(h) \tag{16}$$

9 Indeed, in this case, the partial information provided by h reduces to the categorical one (see definition 1). However, according to definition (11), h provides some partial information about $\pm a_i$ also when h entails neither a_i nor $\neg a_i$.

Second, one can again obtain from (16) a definition of verisimilitude simply by changing the sign in the right side of the equation:

$$vs(h) \equiv inf_T(h) - inf_F(h) \tag{17}$$

In words, the verisimilitude of h is the difference between the amounts of partial true and false information provided by h.[10] The following two equalities are the counterparts of equations (9) and (10):

$$vs(f) = -1 \leq vs(h) \leq 1 = vs(t) \tag{18}$$

$$vs(\top) = 0 \tag{19}$$

As before, these equalities mean that $vs(h)$ varies between a maximum given by the verisimilitude of the truth and a minimum given by the verisimilitude of its specular. Moreover, the verisimilitude of a tautology discriminates between those h for which the amount of partial true information exceeds the amount of partial false information, and hence $vs(h) > 0$, and those for which the opposite is true, and hence: $vs(h) < 0$.

Table 1 displays the degrees of verisimilitude of the 16 propositions of our toy example in Figure 1, as measured by both Vs and vs (and assuming that $p \wedge q$ is the truth in L_2).[11] Note that the two measures agree on all the conjunctive

Table 1. The verisimilitude of the 16 (logically distinct) propositions of L_2 as assessed by measures Vs and vs, assuming that $p \wedge q$ is the truth. True propositions are on the left, their false negations on the right.

	h	$Vs(h)$	$vs(h)$		$\neg h$	$Vs(\neg h)$	$vs(\neg h)$
1	$\top$	0	0	9	$\bot$	0	0
2	$p \vee q$	0	0.33	10	$\neg p \wedge \neg q$	−1	−1
3	$p \vee \neg q$	0	0	11	$\neg p \wedge q$	0	0
4	p	0.5	0.5	12	$\neg p$	−0.5	−0.5
5	$\neg p \vee q$	0	0	13	$p \wedge \neg q$	0	0
6	q	0.5	0.5	14	$\neg q$	−0.5	−0.5
7	$p \rightarrow q$	0	0	15	$\neg p \vee \neg q$	0	0
8	$p \wedge q$	1	1	16	$\neg p \vee \neg q$	0	−0.33

10 As an immediate consequence of equalities (16) and (17), we find again that $vs(h)$ can be expressed in terms of $inf(h)$ and $inf_F(h)$ or, alternatively, in terms of $inf(h)$ and $inf_T(h)$: $vs(h) = inf(h) - 2inf_F(h) = 2inf_T - inf(h)$.

11 For completeness, table 1 also includes (in cell 9) the logically false statement $\bot$. According to Popper (1963b, Addendum 3, pp. 393 ff); see also Schurz and Weingartner (2010, sect. 2)),

theories, including the constituents. However, they disagree on the remaining, non-conjunctive theories, since *vs* is, as desired, more fine-grained than *Vs*. To be sure, in such a simple language as L_2, this is evident only for statements $p \vee q$ and $\neg p \vee \neg q$ (corresponding to cells 2 and 16 in the table).

By considering slightly more complex languages, however, it becomes clearer that *vs* is actually an adequate measure of truthlikeness for non-conjunctive theories. A couple of examples will illustrate this point. Assume that $t = a_1 \wedge a_2 \wedge a_3$ is the truth in L_3. Then, as far as disjunctions of basic propositions are concerned, one can easily check that, for instance, the following statements are in increasing order of truthlikeness:

(20)
$$vs(\neg a_1 \vee \neg a_2) \cong -0.22$$
$$vs(\neg a_1 \vee \neg a_2 \vee \neg a_3) \cong -0.14$$
$$vs(\neg a_1 \vee \neg a_2 \vee a_3) \cong -0.05$$
$$vs(\neg a_1 \vee a_2) = 0$$
$$vs(\neg a_1 \vee a_2 \vee a_3) \cong 0.05$$
$$vs(a_1 \vee a_2 \vee a_3) \cong 0.14$$
$$vs(a_1 \vee a_2) \cong 0.22$$

Note that, on the contrary, *Vs* is 0 for all the above statements. The reason, as we said, is that *vs*, but not *Vs*, is sensitive also to small amounts of information provided by weak hypotheses on the basic features of the world. For instance, while *Vs* cannot discriminate between the verisimilitude of $h = a_1$ and that of the slightly stronger hypothesis $g = a_1 \wedge (a_2 \wedge a_3)$ – since one can check that $Vs(h) = Vs(g) = 0.5$, one instead obtains that $vs(g) \cong 0.55 > 0.5 = vs(h)$, i.e., that vs is sensitive to the small increase of true partial information provided by g over h.

---

contradictions should have the minimum degree of verisimilitude; measure *vs* instead assigns them an intermediate degree of verisimilitude (i.e., 0), on a par with tautologies. This is because both tautologies and contradictions, for different reasons, don't really provide any information about the truth: the former being completely uninformative, the latter providing exactly the same (maximal) amount of true and false information. Here, we decided not to consider contradictions as really relevant hypotheses, in line with much discussion on verisimilitude (cf., e.g., Niiniluoto (1987, p. 150)).

4 Concluding remarks

We conclude our discussion by pointing out some general properties of measure *vs*. As mentioned in Section 2, all post-Popperian accounts of verisimilitude eschew what Oddie (2013, p. 1652) calls "the relative trivialization of verisimilitude for falsehoods", which, as the Tichý-Miller theorem showed, plagued Popper's original definition. Our account is no exception, since it is easy to check that:

$$\text{if } h \text{ and } g \text{ are false, it may be that } vs(h) > vs(g)$$

An example is provided by the first two equalities in (20), which show that $vs(\neg a_1 \vee \neg a_2) \equiv -0.22 < -0.14 \equiv vs(\neg a_1 \vee \neg a_2 \vee \neg a_3)$. More interestingly, *vs* also avoids "the absolute trivialization of verisimilitude for falsehoods" (Oddie, 2013, p. 1652), i.e., it meets the following condition:

$$\text{if } h \text{ is false and } g \text{ is true, it may be that } vs(h) > vs(g).$$

As an example, $vs(a_1 \wedge a_2 \wedge \neg a_3) \equiv 0.33 > vs(\neg a_1 \vee a_2) = 0$. Another attractive aspect of *vs* is that it does not satisfy the implausible condition that verisimilitude increases with logical strength among falsehoods:

$$\text{if } h \text{ and } g \text{ are false, and } h \models g, \text{ it may be that } vs(h) < vs(g).$$

The first example given above illustrates also this point: $\neg a_1 \vee \neg a_2$ is stronger, but less verisimilar, than $\neg a_1 \vee \neg a_2 \vee \neg a_3$. More generally, as we said, f is the least verisimilar statement of the language and, at the same time, it is as strong as any other falsehoods can be. In particular, f is stronger but less verisimilar than the negation of the truth, i.e., of the false content element of the language: in our example in L_3, we have that $vs(f) = vs(\neg a_1 \wedge \neg a_2 \wedge \neg a_3) = -1 < -0.14 \equiv vs(\neg a_1 \wedge \neg a_2 \wedge \neg a_3) = vs(\neg t)$.

Finally, and quite surprisingly, *vs* violates the weak value of content for truths, i.e., the condition, discussed in Section 2.3 above, according to which verisimilitude increases with logical strength among truths. In fact, it is easy to check that:

$$\text{if } h \text{ and } g \text{ are true, and } h \models g, \text{ it may be that } vs(h) < vs(g).$$

An example is again provided by the equalities in (20): $vs(\neg a_1 \vee a_2) = 0 < 0.05 \equiv vs(\neg a_1 \vee a_2 \vee a_3)$. The circumstance that *vs* violates the weak value of content for truths has several interesting implications that, due to space limitations,

we cannot explore here. Following Popper (1963b), many other theorists regard this condition as an important desideratum for verisimilitude; see, for instance, Niiniluoto (1987, p. 133) and Schurz and Weingartner (1987, p. 49), Schurz and Weingartner (2010, p. 417). Still, this desideratum is violated by the well-known Tichý-Oddie "average" measure of verisimilitude, and indeed by all "likeness" accounts as characterized by Oddie (2013, p. 1668 ff.). Here we can only anticipate that this striking similarity between *vs* and the average measure is by no means a coincidence, since, in spite of their completely different conceptual foundations, these two measures of verisimilitude are indeed identical.[12]

12 For a proof of this claim see Cevolani and Festa (unpublished).

Antonella Corradini

Mental Causation and Nonreductive Physicalism, an Unhappy Marriage?

Abstract: Peter Menzies is among those contemporary philosophers of mind who have tried most deliberately to make mental causation compatible with nonreductive physicalism, thus proving the invalidity of Kim's causal exclusion argument (Kim (2005), p. 17). The compatibility between mental causation and nonreductive physicalism will be the focus of this essay. In the first part, I shall expound the tenets of Menzies' theory of mental causation. In the second, I shall emphasise the difficulties his theory encounters, that jeopardise his attempt to reconcile mental causation with physicalism, even though the sort of physicalism he champions takes a quite liberal shape.

1 Exposition of Menzies' theory

Peter Menzies is among those contemporary philosophers of mind who have tried most deliberately to make mental causation compatible with nonreductive physicalism, thus proving the invalidity of Kim's causal exclusion argument (Kim (2005) p. 17). Menzies' attempt rests mainly on the distinction between a conception of cause as sufficient cause – where causation is understood as causal sufficiency – and a notion of cause as difference-making cause – where causation is interpreted instead as counterfactual dependence. On Menzies' view, "the fundamental error of this principle (that of exclusion) is that it mistakes causal sufficiency for causation." (2013, p. 71). Let us now look in more detail at what Menzies means by difference-making cause.

1.1 Truth conditions for making a difference (causal relevance)

> The state S_1 makes a difference to the state S_2 in the actual world just in case 1. if in any relevantly similar possible situation S_1 holds, S_2 also holds; and 2. if in any relevantly similar situation world S_1 does not hold, S_2 does not hold (2013, p. 73).

To Sergio, for 35 years of lively discussions, philosophical and not.

Antonella Corradini: Catholic University, Milan

DOI: 10.1515/9783110529494-007

More formally, the definition of difference-making cause is introduced through the semantics of counterfactuals:

> S_1 makes a difference to the state S_2 iff 1. $S_1 \Box\!\!\rightarrow S_2$ and 2. $\neg S_1 \Box\!\!\rightarrow \neg S_2$ (2013, p. 74).

Menzies' interpretation of counterfactuals is borrowed from the standard possible- worlds semantics of David Lewis (1973). According to Lewis, truth conditions for counterfactuals should be expressed in terms of a similarity – or closeness – relation between possible worlds. The similarity relation "is represented by an assignment to each possible world *w* of a system of spheres of worlds centred on *w*. The system of spheres conveys information about the similarity of worlds to the world *w* at the centre. The smaller a sphere, the more similar to *w* are the worlds in it. So whenever one world lies in some sphere around *w* and another lies outside it, the first world is more similar to w than the second. In terms of this system of spheres, (...) the truth conditions for counterfactuals are as follows: $P \Box\!\!\rightarrow Q$ is true in world *w* if and only if *Q* is true in all the closest *P*-worlds to *w*." (1973, p. 74).

It is worth noting that, in the semantics that Menzies elaborated with List in their (2009) essay, the smallest sphere containing the actual world w always contains more worlds besides *w* – this differs from Lewis' original semantics. The reason for this difference rests on the fact that conditions 1. and 2. must be able to exclude difference-making causes that are either too specific, or not specific enough. The following example, devised by Yablo (1992) and taken up by Menzies, will illustrate both the notion of difference-making cause, and the specific characteristic of the aforementioned List-Menzies' semantics. The example will be developed further in the second part of this essay, in reference to aspects that Menzies himself considers as problematic, and those he neglects by considering them to be irrelevant.

1.2 Yablo's example of the pigeon, and its treatment in the semantics of counterfactuals

Yablo gives the example of a pigeon (Sophie) that is trained to peck a red target when such a target is shown to it (p. 257ff). On Menzies' counterfactual approach, the following counterfactuals thus hold true:

1 Target is red $\Box\!\!\rightarrow$ pigeon pecks
2 Target is not red $\Box\!\!\rightarrow$ pigeon does not peck

Clearly, the truth of the two counterfactuals attests that the target's redness is the cause of the pigeon's pecking. On the contrary, even if it is true that

1' Target is scarlet $\Box\!\!\rightarrow$ pigeon pecks

we cannot affirm that the target's being scarlet is a relevant cause of the pigeon's pecking, because the following counterfactual does not hold true:

2' Target is red, but not scarlet $\Box\!\!\rightarrow$ pigeon does not peck

In fact, even though the target is not scarlet, it is true that the pigeon pecks it, since the target is still red, albeit a different hue from scarlet.

Placing this example in the context of List-Menzies' semantics permits us to capture all of its clarifying power. To say that the red is the cause of pecking means that in all of the closest worlds to the actual world *w*, where it holds true that the target is red, it also holds true that the pigeon pecks. Now, among those worlds there are some whose hue of red is different from that of the target in the actual world. For example, there is a scarlet target. Thus, since the pigeon pecks in those worlds, too, the counterfactual 1' is valid. But among these worlds, there are also ones whose target is red but not scarlet. Despite that, the pigeon pecks in those worlds, too, so that the counterfactual 2' is falsified. Conversely, let us suppose that the pigeon is put in front of a target that is colourful, but not red. Clearly, if the relevant conditions are always the same, the pigeon does not peck. In this case, being the first counterfactual falsified, one cannot say that being colourful is the relevant cause of pecking.

What is the meaning of all this? Well, List-Menzies' semantics allows us to identify the actually relevant causes, i.e. those that are neither too specific nor not specific enough. In effect, taking into consideration the many worlds where the various shades of red may occur, allows us to highlight the actual relevant cause of pecking. This does not consist in a specific hue of red, but in its being abstractly red. In other words, the requirement that the closest worlds to *w* are many, avoids the danger of regarding the relevant cause to be the possession of too specific a quality. Clearly, if the set of the closest worlds to *w* were the singleton of *w*, the relevant cause of pecking would be some more specific quality than merely being red. By contrast, that the pigeon does not peck if the target is merely coloured means we can exclude the notion that the relevant cause could be less specific than red.

From the example of the pigeon, we can now move to the analysis of the general structure of mental causation.

1.3 Counterfactual analysis of mental causation

On List-Menzies' theory of relevant causation, mental event M is the cause of B – the rising of the arm – if the following two counterfactuals hold true:

1. $M \Box\!\!\rightarrow B$
2. $\neg M \Box\!\!\rightarrow \neg B$

Between M and B there is the same kind of relationship that exists between the red target and the pecking. As with the case of the pigeon, the relevant cause must not be overly specific, so the mental cause must not be so specific to encompass its realiser – say N_i. In this case, in fact, B's relevant cause would be no longer M – in its multiple realisability – but M, as it is instantiated by its specific realiser N_i. In other words, the subject of the arm's rising would not act as moved by the mental state M, but as realized by N_i.

The symmetry between the two cases is also reflected in the fact that M and its realiser N_i are not counterfactually parallel. Indeed, although

1' $N_i \Box\!\!\rightarrow B$

holds true,

2' $\neg N_i \Box\!\!\rightarrow \neg B$

does not, because it is possible that another realiser of M – say N_j – causes B, the arm rising. The language of counterfactual semantics is very enlightening on this point: in the closest B-worlds, where N_i does not hold true, it is possible that M is realized by N_j.

Menzies' main purpose in conducting the previous analysis is to revise the exclusion principle. However, in this essay, I shall not follow in detail Menzies' criticism of the traditional exclusion principle. Instead, I shall focus on the revised principle, as it is presented in List and Menzies (2009), and in Menzies (2013) and Menzies (2015).

1.4 Revised exclusion principle

According to List-Menzies, the exclusion principle features two formulations: the upwards and the downwards.

a. Revised exclusion principle (upwards formulation): if state S causes state B, then no state $S*$ that supervenes on S causes B.

In its application to the mental domain, the principle establishes that the subvenient cause – the neural event N_i, that is, the realiser of mental event M – excludes the supervenient: the mental state M.

b. Revised exclusion principle (downwards formulation): if state S causes state B, then no state $S*$ that realizes S causes B.

In its application to the mental domain, the principle establishes that the supervenient cause – the mental event M – excludes the subvenient: the realiser N_i of mental event M.

It is easy to see that the upwards exclusion principle coincides – in its results – with Kim's exclusion principle, whilst the downwards principle seems to be an open denial of that. This is because of the difference between the notion of cause displayed in Kim's original principle and that displayed in Menzies' new version. According to Kim, the cause is conceived as a sufficient cause, whereas, according to Menzies, the cause is to be understood as a relevant cause. This allows us better to grasp why the relevant cause of B is M, and not its realiser N_i: M could be realised by realiser N_j, instead of N_i.

However, Menzies is quite cautious in introducing principle b. In fact, he adds the concept of sensitivity of a mental cause M to a physical realiser N. By deploying this notion he obtains a more refined version of the downwards exclusion principle. Let us now examine how the sensitivity concept looks.

The causal relation between M and B is realisation-sensitive if B does not hold true in all the closest $\neg N$-worlds to the M-worlds (that is, worlds where M has a different realiser than N, that is, the actual one).

Conversely:

The causal relation between M and B is realisation-insensitive if B holds true at least in some of the closest $\neg N$-worlds to the M-worlds (that is, worlds where M has a different realiser than N, i.e. the actual one).

Now, depending on the definitions introduced, Menzies can obtain two remarkable results.

1 Compatibility result: if M causes B, then N causes B iff the causal relation between M and B is realisation-sensitive.

And

2 Downwards exclusion result: if M causes B, then N does not cause B iff the causal relation between M and B is realisation-insensitive.

Although these results are a reformulation of the revised exclusion principle, they are nevertheless significant, since they allow us to derive some important implications. The first is the revision to which Menzies submits the principle of causal closure, and the criticism he levels at it. Further implications concern the specific nonreductive physicalist solution to the problem of mental causation that is endorsed by authors with whom Menzies can be compared.

1.5 Revised principle of causal closure

In the light of his theory of causation, Menzies not only revises the exclusion principle, but also the principle of causal closure. The latter is usually formalised as follows:

CCP Every physical effect has a physical sufficient cause (2015, p. 23)

So far, the direction in which Menzies imparts his revision of the principle should be clear. Since he sees the cause as relevant cause and not as sufficient cause, the principle must be changed accordingly. The notion of sufficient cause must be replaced with that of relevant cause. As a result, we obtain the following principle:

CCPR For every physical effect there is a physical difference-making cause (2015, p. 37)

Revised in this way, it is clear why Menzies criticises the principle: he maintains that it is false. If the mental causes are not physical, and the mental cause – under normal circumstances – is the relevant cause, there cannot – on pain of overdetermination – be a relevant physical cause of the same effect.

1.6 Some implications concerning the nonreductive physicalist solution to the problem of mental causation

Several nonreductive physicalists support a compatibility argument on which physical behaviour *B* follows causally both from *M* and its physical realiser *N*.

Thesis of compatibility: "Any piece of intentional behavior has two causes: a mental state and the neural state that realizes it" (2013, p. 81). The mental and the neural cause are not partial causes, in the way that the presence of oxygen

and a short circuit are each partial causes of a fire. Indeed, on the nonreductive physicalist approach, the mental and the neural state are each causally sufficient for the physical effect. The overdetermination of the effect on the part of the two causes, moreover, is not problematic, since the causes are not accidentally overdetermined – as in the case of the two assassins who kill the same person – but, rather, are essentially overdetermined, in so far as a necessity relation occurs between the mental cause and its realiser.

Among supporters of the compatibility thesis, Menzies mentions Shoemaker (2007) in particular. It will be useful to afford brief attention to Shoemaker, because he deploys Yablo's relation between determinable and determinate in order to illustrate the relation between the mental cause and the neural.

Shoemaker, referring to the example of pigeon, Sophie, maintains that the scarlet (the determined) is a realiser of the red (the determinable), because the causal powers of the red are a subset of the causal powers of the scarlet. Here, saying that an instantiation of the red is a cause of the pecking, is fully compatible with saying that an instantiation of the scarlet is also a cause of the pecking. We just need to add that the realiser is a cause, because, even though only the causal powers of the red are necessary for Sophie's pecking, they are inherited from the scarlet. In other words, the instance of the scarlet is a sufficient cause of Sophie's pecking, but only as far as it coincides with an instance of the red (2007, p. 14).

Overdetermination is avoided, because – although the mental state M and the neural N are both causes of behaviour B, N has its causal role in virtue of the causal role of M. Shoemaker gives the example of Smith's death by a salvo of shots from a firing squad, of which only one – the one shot by Jones – hits Smith. In that case, it can be said that the salvo of shots fired by the whole squad causes Smith's death, in virtue of the shot fired by Jones.

However, Menzies continues to be critical of the compatibilist thesis. He questions (2013, p. 82) the two principles on which the thesis rests: the idea of causation as causal sufficiency, and the principle of transmission of causal sufficiency through the realization relation. Menzies puts forward several reasons in order to rebut these principles. Yet, his main reason derives from the downwards exclusion principle. The consequence to be drawn from this principle is that, when a mental state M causes a behaviour B in a realisation-insensitive way, then the neural state N that realises the mental cannot be the cause of B. Thus, under realisation-insensitivity, downwards exclusion rules out the double cause required by compatibilism. This result is reinforced by Menzies' conviction that cases of realisation-insensitivity are normal, while those of realisation-sensitivity represent an exception.

2 Physicalism, causal closure, downwards exclusion principle

2.1 Determinable/determinate, mental cause/physical realiser

Menzies uses the example of Yablo's pigeon to illustrate the difference between sufficient cause and relevant cause, and to show reasons for preferring the relation of causation in terms of relevant cause to that of sufficient. In Yablo's example, moreover, the relation between determinable (determinable property) and determinate (determinate property) plays a pivotal role with respect to the relation between relevant cause and its realiser. Menzies, instead, is quite sceptical about this parallelism. "The account of the causal efficacy of mental properties that Yablo provides is, unfortunately, limited in its application because it presupposes that mental properties are related to their underlying neural properties as determinables to determinates. In its place, I shall offer an alternative account of causal claims" (2008, p. 197). In its place, Menzies proposes an alternative conception of the cause – one that underlines the contrastive character of causal relations. Nevertheless, although he is critical of Yablo's approach, he accepts his basic tenets. For this reason, Menzies' theory, like Yablo's, should be subject to criticism.

In the first place, Menzies neglects an aspect of the relation between determinable and determinate that, in my opinion, is important for the very distinction between sufficient and relevant cause. What is the aspect he neglects to consider? He disregards the significance of the causal relation as a relation of production (on this, see Crane (2008), par. 5; Hall (2004)). This prevents him from seeing that only the determined, not the determinable, can play the role of a cause as a productive, hence sufficient, cause. In fact, the determinable relates to the determinate as an abstract to the concrete. The determinable is abstract in the sense of being a universal with respect to the determinate, which is a particular. Now, only the concretes (objects or states of affairs) can be causally efficacious. That is why the causation relation cannot exist between an abstract and a concrete behaviour, but only among concretes – i.e. among concrete events or facts, and among instances of properties and not among properties. If, at this point, we think of red in general as an abstract, and of *this* red, as a concrete, we must admit that it is not abstract red that is causally efficacious, but *this* red, endowed with a specific shade, hue or brightness. In conclusion, within the conception of cause as production capacity, it is not abstract red, but only *this* red that can be a cause, since it does not make sense to say that to be abstractly *A* causally implies to be *B*; at most, we can say that *x as A* causes *B*.

This analysis – in appealing to the causal relation as production – does not directly touch Menzies' theory of relevant causation. From the causal point of view, even an abstract aspect can be relevant, like an abstract property that is instantiated in different ways. Nevertheless, the phenomenon of causation would not take place if it did not occur among concrete events. This means that the relevant cause could not play its difference-making role if it did not act within a complex of conditions that enabled it to produce the effect. How should this complex look? It should be such as to guarantee that the additional conditions it encompasses – when united with the relevant cause – are able to product the effect. But this could not occur if the state of affair constituted by these conditions, plus the relevant cause, were not maximally determined. That is, the complex must be a concrete state of affairs, since only a concrete state of affairs, as previously said, is endowed with causal efficacy. Thus, in order for M to be the relevant cause of behaviour B, it is necessary that there is a set X of conditions such that $X \wedge M$ is a sufficient condition of B. Of course, X should also encompass a concrete physical realiser N of M. In conclusion, just as the properties of the objects involved in an event could not be relevant causes of the event-effect if there were not the entire sufficient cause of the event-effect, so M could not be the relevant cause of B if $X \wedge M$ were not a sufficient condition of B.

If the notion of sufficient cause –, even when distinct from that of relevant cause –, is presupposed by the latter, why does Menzies not take this into account in the formulation of the causal closure principle, instead of rebutting its usual formulation? Note that this question appears still more pertinent if we consider what Menzies himself says –, that is, that the weakening of the centring requirement is owing to the fact that the counterfactual $P \mathrel{\square\!\!\rightarrow} Q$ must "express the condition that P would be sufficient in the circumstances for Q" (2015, p. 31). In my opinion, the only reason that Menzies could give was that he was only ready to assume one form of causal relation: that expressed through relevance. The notion of a cause is unique because that of sufficient condition (of course, without the concept of production cause) is absorbed into the relevant cause. This is also why Menzies reformulates the causal closure principle in terms of difference-making cause, and, for this reason, rejects it. But this move is not justified, if the concept of sufficient cause – albeit absorbed into the relevant cause – continues to be presupposed.

2.2 Downwards exclusion principle: some open questions

From downwards exclusion principle one obtains that, if M causes B and the causal relation between M and B is realisation-insensitive, then N is not cause

of B. This result poses a number of problems for those who, like Menzies, support nonreductive physicalism. Indeed, what does it mean that the causal relation between M and B is realisation-insensitive? In List and Menzies (2009) we can find the following definition: "Call the causal relation between M and B realization-insensitive if B continues to be present even under some small perturbations in the realization of M, or formally, if B is present in some closest $\neg N$ worlds that are M-worlds" (p. 496). A better definition is proposed in Menzies (2015): "An event M that is actually realized by a physical event N is realization-insensitive cause of another event P (with respect to N) iff a) M is a difference-making cause of P; and b) some of the closest $\neg N$ world in which M holds are worlds in which P holds", (p. 40).

Let us try to comprehend the second definition accurately. If we stick to its formal meaning, the definition reads that M is realisation-insensitive with respect to its actual realiser N, if a) B is present in all the closest M-worlds to the actual world and b) in some of those worlds N is not present. Now one wonders, in the M-worlds where B occurs, but realiser N is not present, is any other physical realiser present, which, say, plays the role of the actual realiser N? On the basis of the first definition, it seems that the answer is positive, since the authors speak of "small perturbations in the realization of M"; however, the formal structure of the definition does not exclude more daring kinds of interpretation. I believe that the preferable interpretation depends on the concrete meaning one is willing to give to the relation of closeness. Now, if the nonreductive physicalist who shares Menzies' causal notion wants to keep his feet firmly on physical terrain, he cannot accept that the causal relation between M and B is insensitive with regard to all the physical realisations taken collectively, meaning that M might be a difference-making cause of B, regardless of any realiser. The nonreductive physicalist can only accept that insensitivity regards each of the realisers taken one by one, so that, if realiser N_i does not exist, the alternative physical realiser N_j does exist. But if this is so, it means that the relevant cause of B is not M, but the existence of a realiser of M. Indicating with *phyreal x* of M the relation of being a physical realiser of M, it would, in fact, be true that:

1 $\exists x$ (*phyreal x* of M) $\Box\!\!\rightarrow B$
2 $\neg\exists x$ (*phyreal x* of M) $\Box\!\!\rightarrow \neg B$

Following this result, what is the role of M? It does not cease to be relevant, but the analysis reveals M's nature as a formal cause more than an efficient cause. M is a formal cause because it is a structural component common to all its realisers. On Menzies' approach, moreover, it is uncertain whether M is endowed with causal powers that are not inherited from its neural basis. The most that can be said,

from within a physicalist framework, is that the cause of B is a certain N_i as realiser of M.

2.3 A dualistic outcome? It seems not

At this point, one can see the convergence of the results obtained in the previous two paragraphs. The common outcome is that the mental cause could not have causal relevance if behaviour B were not produced by the specific realiser that, in turn, plays the role of instantiating the cause. Just as the red is the difference-making cause of Sophie's pecking because the target's concrete colour causes the pigeon's pecking *as far as* it is red (and not, for example, *as far as* it is scarlet), so M is the difference-making cause of B because the cerebral event N is the sufficient condition of B *as far as* it is a realiser of M (and not of some other mental event).

However, in the examination of mental causation, we cannot stop at this analogy. It is hardly possible not to be doubtful about the hypothesis that realiser N derives from a mere concretisation of the mental event. Mental events are not abstract entities, like red (as universal) with respect to *this* red. My current thought that "now it is raining" is a concrete mental event, localised in time, and attributable to a specific subject. But, then, that which is concrete cannot be made even more concrete – or, that which is superdetermined cannot be further determined. Therefore, the relation between M and N cannot be equated to that between universal and concrete.

That the mental cause is not related to its realiser as an abstract to a concrete, but as something concrete to something else concrete, should make us think. Strictly speaking, this asymmetry does not allow us to say that M causes B, because the property that competes to N is not M, but that of being a *realiser* of M. So, if one keeps supporting the thesis of the sufficient cause as concrete production capacity, it should be said that the relevant cause (in counterfactual sense) is not M, but *being a realiser* of M: not *this* realiser, but any of its realisers. Menzies, of course, would non accept this conclusion, since he would reject the notion of sufficient causation. But if we take into account the reflexions we have developed in paragraph 2.1, we cannot exempt ourselves from accepting the virtuous compatibility of the two theories of causation. This leads us straight towards a *contributing* conception of mental causation: the mental cause and some of its physical realisers are *essential*, thus *both relevant*, *components* of the sufficient cause of behaviour. In the subsequent pages, I want to discuss this view in the light of the critical remarks that Menzies makes in (2015), pp. 38–39.

2.4 Breaking of causal closure and contributing theory of causation

As we have already mentioned, in (2015), Menzies intends to show that the causal closure principle, formulated in terms of relevant cause, is false. On his view, it is false that "for every physical effect there is a physical difference-making cause" (p. 37). It is false because the relevant cause is the mental, not its physical realiser.

The rejection of this principle may seem an element that could rupture mental causation's physicalist character. Indeed, some authors have wondered whether the refusal of this principle is compatible with Menzies's explicit endorsement of physicalism, even when understood in a nonreductive way (Bermudez and Cahen (2015), p. 54). As I have argued in 2.2., it seems to me that Menzies' rebuttal of causal closure is not incompatible with physicalism. There is no question of incompatibility, if, although M's relevance consists in its insensitivity to the single realiser, this does not rule out its dependence on any realiser in each of the *M*-worlds where *B* is present. In fact, the falsity of the revised closure principle does not regard *M*'s being nevertheless physically realized. But, then, the breaking of the causal closure – as hypothesised by Menzies – would not be as significant as it might seem at first glance. Only if the mental cause were independent from its realisers in a stronger way – so to be able to exert *real* causal powers besides those inherited by its realisers – would the breaking of the causal closure principle be *real* indeed. In that case – that is, if the mental cause had real causal powers that were not not inherited from its physical realisers – the mental event should be understood as an essential contributing cause, along with the (likewise essential) physical contributing cause, consisting in its actual realiser. In this way, the relation between mental and corresponding neural event would not be between cause and its realiser, but between two events that contribute together to the realization of the effect.

The plausibility of this interpretation is confirmed by the fact that it is briefly discussed and rejected by Menzies in paragraph 3 of (2015). According to him, we cannot treat the realiser *N* the same way as the mental cause, since: "... the mental event *M*, differently realized, would have produced the same effect. So in the light of this, it is difficult to maintain that *M* and *N* should be assimilated to the paradigm of contributing causes" (p. 39). However, this criticism is not compelling. First of all, we can reply that *M* is an essential contributing cause together with any of the realisers that – although not essential if taken *singularly* – are essential if taken *disjunctively*. Secondly, the contributing thesis is necessary if one maintains that *M* is bearer of causal powers that are not inherited by *N*. To

express the idea that the mental cause is endowed with its own causal powers, the relation of closeness among worlds can be defined as follows. Let

$$Z \wedge N \wedge M \Rightarrow B$$

express the relation of sufficient causation of B by $N \wedge M$ in the context of the remaining conditions Z. Let Z contain all that is relevant in order for B to occur in virtue of $N \wedge M$ – for example, let Z contain the set of the natural laws in force in the actual world, but also the exclusion of M's realisers that are actually not active. Thus let the closeness relation among worlds be determined by Z. Then

1 $N \wedge M \ \Box\!\!\rightarrow B$ (in all the closest $N \wedge M$-worlds to the actual world B holds);
2 $\neg(N \wedge M) \ \Box\!\!\rightarrow \neg B$ (in all of the closest worlds to the actual world where $N \wedge M$ does not occur, B does not hold).

This means that the difference-making cause of B in the actual world is a composite cause, whose components N and M are both essential (on the notion of composite cause, see Corradini (2015), p. 117).

3 Conclusions

Nonreductive physicalism – in any of its versions – needs a robust notion of mental causation. This is necessary to distinguish it from forms of reductive physicalism, like identity theory, that gives special importance to the physical realisers of mental states. A decisive aspect of nonreductive physicalism is that the mental cause should be – to a certain degree – independent from its physical realisers. But when is a mental cause independent from its realisers? Menzies explains that this is the case when it meets the requisite of relevance. However, although Menzies' theory of causal relevance is sound and accurately designed, some of its aspects threaten its robustness.

First, the difference-making cause is more similar to a formal cause than to an efficient one; moreover, its relevance lies on the notion of necessary cause, rather than sufficient and production cause. Furthermore, the adoption of Yablo' s parallelism between determinable/determinate and mental/neural only reinforces the mental cause's flaw of being merely a formal cause.

But all of these aspects signify a larger problem that affects physicalism: be it reductive or nonreductive, the independence it assures to mental cau-

sation pertains, at most, acknowledgment of its relevance. It does not go far enough to afford it autonomous causal powers that are not derived from their physical realisers. As a consequence, not even Menzies – as a nonreductive physicalist – has succeeded in finding the right place for mental causation within physicalism.

Ciro De Florio

On Grounding Arithmetic

Abstract: Philosophy of mathematics of last fifty years has been dominated by the metaontological stance according to which one fundamental problem of the ontology of mathematical theories is the existence of mathematical objects and the related epistemic access to them. But during the last ten years another fecund and promising metaphysical framework has been developed: the key idea (which goes back to Aristotle) is that the main problem of metaphysics is about the relation of grounding among various levels of reality. Although this approach should be relevant for almost all the metaphysical questions, however, there are few attempts to extend these intuitions to the debate in philosophy of mathematics. The aim of this, preliminary, work is analysing some possible outcomes of the grounding approach in metaphysics of mathematics.[1]

1 Easy proof of existence of numbers

Let us take into exam the following inference:

(P) There exist prime numbers
(C) There exist numbers

It is an one-premise argument which appears to be perfectly valid: if the premise (P) is true, it will be true the conclusion (C) too. Obviously, the problem is whether the inference is also sound, namely, if the premise is true. However, it is a trivial truth of naïve arithmetic that there exist numbers divisible just for 1 and for itself. But then, the conclusion follows: there exist numbers. It would seem that oceans of ink (since Plato) has been wasted: mathematics does not push in front of a complicate ontological question, the existence of numbers is a trivial

1 This essay is dedicated to Sergio Galvan who, twenty years ago, showed me astonishing universes and taught me a lot of things. But, above all, he taught me how to learn infinite others. I would like to thank Alessandro Giordani and Ilaria Canavotto for their insightful comments on various drafts of this work; the audience of the International Conference *Truth and Ground* (Ascona, 2015); Andrea Sereni, Alfredo Tomasetta, Luca Zanetti and all attenders to the Grounding Seminar (IUSS, Pavia 2015).

Ciro De Florio: Catholic University of Milan

DOI: 10.1515/9783110529494-008

consequence of a trivial truth. On the other hand, it does not seem reasonable to admit that (P) is false:

> (P) is a mathematical truism. It commands Moorean certainty, as being more credible than any philosopher's argument to the contrary. (Schaffer (2009), p. 357)

Paraphrasing David Lewis, how philosophers could be so intellectually arrogant to put in doubts a so basilar truth based simply on the fact that a consequence of (P) is (C), that there exist numbers? More specifically, it is difficult to think that any argument which tries to show that (P) is false can have a degree of rational warrant higher than naïve arithmetic. However, to accept (C) is not philosophically innocuous: numbers are entities supposedly abstract, not located in space and time; moreover, they lack causal powers. For these reasons, it is so difficult to find a place for those exotic entities in a general naturalistic framework. What is, in other terms, the place of mathematical objects in nature?

There are, roughly, two families of strategies to face this problem: according to first approach, the mathematical language has to be interpreted *at face value*. Quoting a famous passage, the truth conditions of the sentence "There are at least three large cities older than New York" are structurally identical to the truth conditions of "There are at least three perfect numbers greater than 17". According to the other approach, the daily mathematical language has to be paraphrased in such way that it can show its actual ontological commitment.

The classical position in philosophy of mathematics belonging to the first group is the Platonism. Usually, the Platonist accepts with no many problems an inference as those previously discussed: reality is not confined to a space-time region. There exist other regions of being, inhabited by other entities. An almost immediate consequence of this ontological point of view is the problem of epistemic access to these entities: if knowledge is essentially connected to any form of causal interaction and if, by definition, mathematical entities lack these kinds of interactions, how can we know something about them? Even in this case, very roughly, the realist philosopher can decide to enlarge his conception of epistemic access (by introducing some form of intuition) or modify the terms of the problem, trying to show that knowledge of mathematical entities is, actually, mediated by the access to particular states of the world less troublesome from a naturalistic point of view.

Who does not want to embrace the Platonist path, but at the same time wants to maintain the at face value interpretation of mathematical sentences, has, roughly, at disposal two strategies about the inference in exam:

(i) He can try to show that the premise (P) as well as the conclusion (C) are literally false. It is not true that there exist prime numbers. And the reason is that there exist no numbers. The entities postulated by mathematical theories are fictional and, as a consequence, mathematics is literally false. It is a strongly eliminative strategy (or *skeptical*, Fine (2001)) advocated by, for instance, Hartry Field (Field (1980), Field (1989)).

(ii) Second strategy, which we can call *physicalist*, is rather neglected (cf. Bigelow (1988)). It accepts the argument in its original form but it denies the background metaphysical assumption, that is, the abstractness of numbers. In other terms, the idea is trying to show that arithmetic is literally true and perfectly acceptable in a naturalistic framework. Mathematical entities and their properties are, according to this view, citizens of a naturalistic world, as all physical entities.[2]

If one think that the theoretical costs of the options which do not alter the standard semantics of the mathematical language are untenable, he can adopt a different strategy and argue for a paraphrase of (P) and (C). The idea is that (P) does not literally express a truth but it must be relativized, for instance, to a specific theoretical framework. Let us assume, for the sake of simplicity, that when one says that it is a basilar arithmetical truth that there exist prime numbers, *actually*, he is stating that "it is provable, in arithmetic, that there exist prime numbers". So, our inference can be paraphrased, roughly, as:

(P*) In arithmetic, it is provable that there exist prime numbers
(C*) In arithmetic, it is provable that there exist numbers

The argument is valid and (very probably) sound too. It is trivially true that in arithmetic it is provable that there exist prime numbers and, then, conclusion follows. But, say the advocates of this strategy, from that it does not follow any direct commitment to the metaphysical claim (C), that there exist numbers. Here one is only committed to certain facts concerning arithmetic proofs. It is still open, of course, the question about the connection between the truth that it is provable that there are numbers and the fact that, actually, there are numbers. In other terms, this is about a sort of bridge-principle which links, in some way, facts concerning mathematical provability with truth tout court. Obviously,

2 This strategy could refer to Mill's intuitions. Field too, in his fictionalist program, shows how to lay down a physical theory (Newton's physics) with no use of mathematical entities, by substituting real numbers with spatio-temporal points, assumed as concrete entities.

the anti-platonist can argue against the existence of these bridge-principles, saving, then, the moorean trust in naïve arithmetic without buying the expensive Platonist metaphysics.

As discussed in Schaffer (2009), Fine (2001), Correia and Schnieder (2012), the reason of this kind of philosophical reactions has to be looked for in the general conception of metaphysical enterprise, dominating for about four decades: as said before, the task of metaphysics is to determine what there is and the method to lead this inquiry is to focus on the ontological commitment of our best theories (cf. Schaffer (2009), p. 348). The outcome of this stance, which we call Quinean,[3] will be a list, an inventory, and consequently, the image of the world will be *flat*, lacking metaphysical structure. On the contrary, for the grounding theorists – who inspire more or less explicitly to Aristotle – the aim of metaphysics is to explain *the relations of dependence and grounding* among various levels of reality, and, hence, to provide a structured image of the world. The consequences for philosophy of mathematics are relevant and already Aristotle underlined the point:

> It is also true also to say, without qualification, that the objects of mathematics exist, and with the character ascribed to them by mathematicians (*Metaph.* 1077b31-3)

It is obvious that numbers exist; little more than a triviality, given the almost indubitable truth of arithmetic. This does not mean that numbers are not interesting from the philosophical point of view: simply, what is interesting is not their presence in an hypothetical universal catalogue but their *fundamentality*. How do numbers exist? Is their existence independent or are they grounded in more fundamental levels of reality which do not have an arithmetical nature? In Aristotelian terms we could ask: are numbers substances?

The ontology of mathematics is dominated by the Quinean paradigm; for this reason it is interesting to investigate how to apply the intuitions of grounding to these problems. In this preliminary work I would try to apply some intuitions about a metaphysical framework of grounding to problems of philosophy of mathematics. In particular, I shall take into exam two options: to consider arithmetical facts as fundamental and, conversely, to consider them as grounded in other basilar facts. Before that, however, it is worthy to spend some words on the very relation of grounding I am applying in the following. Basing on (Correia

3 That does not mean that all who share Quine's general view about metaphysics agree with his specific thesis; Peter van Inwagen and Stewart Shapiro are two important examples, respectively for general metaphysics and philosophy of mathematics.

and Schnieder (2012), Fine (2012)) let us define the grounding as the fundamental metaphysical tie which links *facts*:

$$[A] \triangleright [B]$$

This can be read as: the fact that A grounds the fact that B; B because (or since) A; B happens in virtue of A and so on. Let us assume that the grounding relation is irreflexive (that is, it does not hold that $[X] \triangleright [X]$), asymmetrical, (that is, if $[A] \triangleright [B]$ then it is not that $[B] \triangleright [A]$), and transitive: $[A] \triangleright [B]$ and $[B] \triangleright [C]$, then $[A] \triangleright [C]$. By embedding questions of philosophy of mathematics in this grounding conceptual framework, immediately the following problem arises: simple arithmetical facts as $[2+3=5]$ seem to be the prototype of necessary facts, and nevertheless "it seems that it's apt for explanation in terms of facts about numbers, mathematical structures, or the like. Indeed, we seem to be possessed of the resources to ground some amongst our necessities." (Bliss, Trogdon 2014)

In other terms, the intuition here at stake is that although the arithmetical facts as very simple equivalences can be rightly considered necessary, nevertheless these facts seem to be dependent on more basic facts.[4]

Enough for the preliminaries; it is time to start our analysis. This paper is essentially divided in two parts which concern, respectively, the position according to which the arithmetical facts are fundamental and the position for which they are not fundamental. As a matter of fact, the latter will be the most rich in philosophical ideas.

2 Fundamental arithmetical facts

Mathematical Platonism advocates the existence and the fundamentality of mathematical, say arithmetical, facts. Historically, it can be traced to Dedekind the most comprehensive attempt to formally characterize the intuition concerning facts about natural numbers. And there is room to claim that Dedekind's axiomatization does catch the structure of natural numbers. That is described by axioms in which just three primitive terms occur: the zero, the succession function and the predicate "being a natural number".

And Dedekind's theorem confirms the success – at least from a certain point of view – of the proposed axiomatization: any system of objects which satisfies

4 In the following, we do not take into account Fine's idea according to which the metaphysical necessity has to be grounded on the essences of things.

the axioms of arithmetic is structurally identical. And in a realistic perspective, the categoricity is surely a relevant result.

However, thing are not so plain. To guarantee the categoricity, the axioms of arithmetic must be expressed in the second order predicate language. This remarkable expressive increase makes the system computationally intractable: second order logic is semantically incomplete.[5] Mainly for that reason, the foundational studies are progressively abandoned the theories expressed in higher order logic. If, on the other hand, one adopts Peano first-order arithmetic, which is semantically complete, he looses the categoricity: there are a (infinite) plethora of *different* systems of objects which make true the axioms. They are the famous non standard models of arithmetic.

Now, which consequences for the view according to which there are fundamental arithmetical facts? If one advocates this position, it is plausible to suppose that there is just one class of genuine arithmetical facts; the other systems of objects – which make true the axioms but are not isomorphic to the standard model – should not be included among the fundamental arithmetical facts.

The reason for that is easy: in the non standard models there exist infinite elements which possess an infinite number of predecessors. Now: either it is a fundamental fact that there are non standard numbers or it is a fundamental fact that there are not. Various reasons suggest to consider the standard model the natural choice. If true, that suggests a possible answer to the question about the identity of fundamental arithmetical facts: they are the series of standard natural numbers, i.e. what Peano-Dedekind axioms characterize up to isomorphism.

3 Non fundamental arithmetical facts

If one does not want to state that there are fundamental arithmetical facts and, at the same time, he aims to save the soundness of very common inferences as the previous, it is natural to consider the existence of arithmetical facts as dependant on other kinds of facts. The question of ontology of mathematics is, therefore, stated as: (natural) numbers and their relations exist, of course, but they are not the fundamental entities of this world. Let us see, as examples, three philosophies of mathematics which can be interpreted in this light: the eliminative modal structuralism, an example of formalism and the neologicism provide respectively

5 But Dedekind cannot know this; it has to wait about fifty years since the publication of *Was Sind und Was Sollen Zahlen* so that Kurt Gödel discovered the incompleteness of second order logic.

the conceptual resources to reduce arithmetical facts to categories of more basilar facts which concern (possible) structures, proofs, and relations among concepts. In the following, we will roughly present the key issues of any proposal; then we will discuss the possible relations between grounding and philosophy of mathematics.

3.1 Structuralism

The key intuition of structuralism (cf. Reck Price, Hellman, Resnik, Shapiro, Benacerraf) is roughly the following: mathematics – in this case, arithmetic – does not talk about objects but about structures, namely complex entities in which what is ontologically relevant are the relations among the items and not the allegedly nature of them. Arithmetic, then, does not describe properties of numbers, intended as individuals, but investigates the features of the structure of natural numbers, in which the single numbers are nothing but place-holders, with no specific property. The number "three" has no essence but being the successor of the successor of the successor of 0.

The eliminative modal version (settled by Hellman) provides a paraphrasing procedure of the sentences of ordinary arithmetic. Let us see how it works.

First of all, let PA^2 be the conjunction of the second order Peano Axioms. The only non logical terms are: $N, s, 0$ which mean respectively "natural number", "successor", and "zero". For convenience, let us indicate the non logical terms occurring in axioms as: $PA^2(0, s, N)$. Now, let us take into exam any sentence of naïve arithmetic as $2 + 3 = 5$. Formally: $+(s(s(0)), s(s(s(0)))) = s(s(s(s(s(0)))))$. For convenience, let us call this sentence: $A(0, s, N)$. Now, we have that the ordinary sentence of arithmetic $2 + 3 = 5$ is equivalent to $PA^2(0, s, N) \rightarrow A(0, s, N)$. But we can further generalize and substitute to every non logical term a suited variable. 0 will be a variable x, the successor function a functional variable f and, in the end, the predicate of being a natural number, a predicate variable X. We have, so, the propositional functions: $PA^2(x, f, X)$ and $A(x, f, X)$. At this point, nothing prevent us to universally quantify:

$$A \leftrightarrow \forall x \forall f \forall X (PA^2(x, f, X) \rightarrow A(x, f, X))$$

Let us notice some things. First of all, any sentence of arithmetic becomes the universalization of a conditional:

> Pure mathematics is the class of all propositions of the form "p implies q", where p and q are propositions containing one or more variables, the same in two propositions, and neither p nor q contains any constants except logical constants (Russell (1903), p. 3)

Secondly, if the paraphrase is sound, there is no more reference to arithmetical states of affairs; in other terms, A does not "actually" speak about numbers but about structural relations among individuals, classes, and functions.

> Instead, what we assert with an arithmetic statement A is now something about all objects, all one-place functions, and all one-place predicates or sets; since the main logical operators in A are unrestricted universal quantifiers. (Reck and Price (2000), p. 356)

When $PA^2(x, f, X)$ is true? When a certain system of objects satisfies the structural properties indicated in the formula; specifically, $PA^2(x, f, X)$ is satisfied by any discrete linear order with a first element and no last element, in which any two elements are separated by a finite number of links. Such structures are normally called ω-sequences. The nature of the items of the ω-sequences as well as of the relations which are defined on them are totally irrelevant; if the succession of Roman emperors had gone on forever, the sequence:

> Augustus, Tiberius, Gaius, ...

would have been another instance of this pattern. The insight of structuralism is that mathematics – in this case arithmetic – does not talk about numbers and relations among them but it refers to any ω-sequence. This form of structuralism proposes to paraphrase the ordinary sentences of arithmetic (as $2 + 3 = 5$) by universal generalizations rather complex. There is the problem of falsity of antecedent, that is, the problem of vacuity of conditional. Russell proposed to assume a sort of axiom of infinity for the lowest type of objects. Less demanding from an ontological point of view is Hellman's modal twist:

$$A \leftrightarrow \Box\forall x\forall f\forall X(PA^2(x, f, X) \rightarrow A(x, f, X))$$

Now, the vacuity can be avoided by a less strong assumption, that is:

$$\Diamond\exists x\exists f\exists X(PA^2(x, f, X)$$

This completes Hellman's modal structuralism: any arithmetical sentence is, actually, a universal conditional which is not about "numbers" or mathematical objects but it is a structural logical relation provided that the existence of an omega sequence is assumed as possible.

3.2 Formalism

Structuralism is not the only proposal of reduction of mathematical truth. According to formalist insight, the objects of mathematics are syntactic patterns of symbols which have not an independent meaning. Mathematical truth is nothing but a form of provability within a suited formal theory. There are (at least) two fundamental issues in formalist (and neo-formalist) program:

i To find a formal theory which grasps the most part of intuitively accepted truths in a specified domain (e.g.: arithmetic, analysis, geometry,...)
ii To warrant the reliability of the chosen theory

Historically, the second point has been the attempt to provide a finitary proof of consistency of formalized theories. Gödel's results ratified sharp limits to this task. Concerning the first point, an interesting choice is first-order Peano arithmetic with the omega-rule. The omega-rule is an inference with the following form:

$$P(0), P(1), P(2), \ldots, P(n), \ldots \vdash \forall n P(n)$$

If a certain property holds of 0, 1, and so on, for any natural number, it holds for all natural numbers. The features of this theory are: PA_ω is not recursively axiomatizable, and for that – one can argue – it is not a completely formal theory. Moreover, PA_ω, if consistent, it is complete, i.e., any arithmetical sentence is provable or refutable within it. For that reason, PA_ω is a good candidate to grasp the intuitive idea of mathematical truth.[6] So, the formalist paraphrase is as the following:

$$T(A) \leftrightarrow PA_\omega \vdash A$$

Therefore, if formalism is sound, we say that is not correct to speak about the existence of numbers and their properties; on the contrary, the only "real things" are determinate syntactical relations of provability.

6 This point is delicate. Of course, there is no agreement about the meaning of "formal" in formalism. If formalism is intended in a quite narrow sense, any attempt to catch mathematical truth by formalism is limited by Gödel's results. So, we can relax these constraints and accept an infinitary theory as Peano arithmetic plus omega-rule. If this choice is acceptable from a formalist point of view is a crucial point but we do not take it into account here.

3.3 Neo-Logicism

Probably, the most animated field of research in philosophy of mathematics during the last three decades is the so-called neologicism (or neo-fregeanism). The neo-logicist insight is close to Frege's original intuition: mathematics is nothing but a branch of logic; therefore, the truths of mathematics are particular kinds of logical truths. The core of the neo-logicist program is to prove (second-order) Peano's Axioms from a strongly enough basic logic plus some principles usually called *principles of abstraction*. Among these, the most known is Hume's principle:

HP $\forall F \forall G(\#F = \#G \leftrightarrow F \approx G)$

here, # is an operator which yields terms starting from concepts and whose intended meaning is "the number of". F and G are higher-order variables which denote concepts whilst $F \approx G$ states that there is a 1-1 correspondence between F and G. The right hand side of the biconditional has two prominent features: it is expressible in pure logical terms (provided that one can quantify on higher order variables) and it is epistemically prior on the left-hand side. (HP) proves the existence of numbers: let us assume that $\#F = \#F \leftrightarrow F \approx F$. Logically, we have $F \approx F$; so, $\#F = \#F$. But then, $\exists x(x = \#F)$, that is there are numbers. A good part of the debate within neologicism takes into exam the ontological inflation of the abstraction principles and their semantic status (for instance, are they analytic?). For our purposes, the key concept is the following: the existence of mathematical objects depends, in a certain way, on specific equivalence relations among concepts.

4 Grounding arithmetical facts

These three examples do not complete, of course, the possible alternatives; but here we are interested in a different question: various philosophies of mathematics suggest different kinds of fundamental facts which supposedly ground the arithmetical facts. It is natural to think that there is just one right answer; therefore all the weight is on the shoulders of the philosopher of mathematics; but how to justify our choice? How to justify that, say, structuralism is the true story about mathematics? There are (at least) two general strategies: according to the first one, a particular option is justified by classical considerations as ontological economy, explanatory power, acceptability from a naturalistic point of view and so on. In

case at discussion, for instance, one could advocate the eliminative structuralism being more sober from an ontological point of view; at the end of the day, all arithmetical facts are grounded, according to eliminative structuralism, just on the possibility of an omega-sequence of elements. And this could be considered a fair prize for the truth of arithmetic.

Alternatively, we can decide to rely on considerations about the grounding relations which would stem from that. For instance, let us assume, again, that we are asking if the modal eliminative structuralism is true; well, we have to look at the grounding relations between "structural" facts and arithmetical facts. According to our schema, indeed, we have that arithmetical facts would be grounded in facts about possible structures. But, one could argue, it is not plausible that facts about *possible* omega-sequences ground *actual* arithmetical facts. Let us notice that the argumentative manoeuvres are different: in the first case, we inquiry the right relation of grounding by means of general considerations which do not entail the very grounding relation between structural and arithmetical facts. In the second case, on the contrary, the true story about the arithmetical facts is told by considerations about the grounding relations.

But what if we were unable to judge what grounding relation is the correct grounding relation? In other terms, what if we cannot know whether the arithmetical facts are grounded in facts about concepts or facts about structures or whatever? This impossibility could be not simply an epistemic limitation; there could be the case that there is no fact of the matter about the grounding relation of mathematical facts and other kinds of more fundamental facts. We will get in a situation like the following:

$[\ldots Structures\ldots] \triangleright [A]$
$[\ldots Proofs\ldots] \triangleright [A]$
$[\ldots Concepts\ldots] \triangleright [A]$

That is, the same arithmetical fact $[A]$ (in our example, the fact that $2 + 3 = 5$) is grounded in very different systems of facts concerning, respectively, (possible) structures, proofs, relations among concepts, and so on. This is a case of over-determination of grounding and, even if it is not inconsistent, there is room for a piece of scepticism.

> It is natural to think these proposals cannot all be correct. If a fact about the numbers obtains in virtue of some fact about the provability of a sentence in PA, it is implausible that it should also obtain in virtue of some quite different fact about all omega sequences (Rosen (2011), p. 129)

How to react to this phenomenon of over-determination?

- This is a case of illusory ground since there is a more fundamental fact, let us call it $[B]$, which is the ground of the facts about structures, proof, concepts and so on. Therefore, the arithmetical fact is really grounded on $[B]$, and the case of over-determination is not problematic. Of course, it still remains a problem: $[A]$ has a plurality of immediate grounds; and one can legitimately ask why $[A]$ is mediately grounded by $[\ldots \text{structures} \ldots]$, $[\ldots \text{proofs} \ldots]$ and so on rather than by $[B]$ directly?
- Since there are different grounds, there are different arithmetical facts. This is Rosen's choice:

> On this view, there is no such thing as the system of natural numbers. There are rather the formalist numbers, facts about which are grounded in facts about provability in PA_{ω}, the modal structuralist numbers, facts about which are grounded in facts about all possible omega systems, and so on. (Rosen (2011), pp. 129–130)

So, our schema becomes:

$[\ldots Structures \ldots] \triangleright [A*]$
$[\ldots Proofs \ldots] \triangleright [A**]$
$[\ldots Concepts \ldots] \triangleright [A***]$

Where $[A*], [A**], [A***], \ldots$ are different arithmetical facts with different grounding relations. The important point is, however, the following:

> Since the differences between these systems of numbers make no mathematical difference, the language and practice of mathematics will have had no occasion to distinguish them. (Rosen (2011), p. 130).

Rosen's idea is to embrace a sort of semantic indeterminateness of standard arithmetic language. Since there are many kinds of arithmetical facts which differentiate just by way in which they are grounded, "it makes no sense to speak of *the* fact that 235 + 657 = 892". As a matter of fact:

> There are rather many equally qualified facts in the vicinity, each concerning numbers of some determinate kind, each grounded in some determinate way in underlying facts (Rosen (2011), p. 130)

Let us briefly discuss Rosen's proposals. We will notice two things: first, if one accepts a relation of truthmaking between facts and propositions, he must accept cases of truthmaking over-determination; second, we will follow Rosen's intuition by showing a possible development.

5 Elaborating

Let us assume that there is a relation of truthmaking between facts and propositions. So we say that the proposition $< A >$ is made true by (or in virtue of) the fact $[T]$:

$$[T] \models < A >$$

So, in our case, we have that the same true arithmetical proposition is made true by different facts:

$$[A*] \models < 2 + 3 = 5 >$$
$$[A**] \models < 2 + 3 = 5 >$$

Now, following Rosen, $[A*]$ and $[A**]$ are different facts, with different grounds. However, they make true the same ordinary arithmetical sentence. The cases of truth-making over-determination are not, as for grounding, incoherent but doubtful. Usually, cases of two (or more) different truthmakers which make true the same proposition are acceptable provided that they are in a way related, say, by an inclusive relation. So, the fact that the ball is scarlet is a truthmaker for the proposition <the ball is red>; but also the fact that the ball is red is a truthmaker for the same proposition. If we assume that they are two different facts we get in a case of over-determination. However, it is arguable that the two facts in question are in a way connected: in particular, one (that the ball is scarlet) is included in a more general fact (that the ball is red). But nothing similar happens in our case: there is no relation of inclusion or the like between facts concerning (possible) structures and facts concerning proofs or relations of equinumerosity among concepts. A very dramatic solution could be to state that there are also *different* mathematical truths but this is not only implausible but also in contradiction with the general assumption of the entire argument, that arithmetic is true. Despite this phenomenon of truthmaking overgeneration, Rosen's proposal is interesting, so let me elaborate his point.

Let $[X*]$ be the class of all arithmetical structural facts (that is, facts which are grounded in the possible structures); $[Y*]$ the class of arithmetical formalist facts (facts grounded in the provability within a suited theory); $[Z*]$ the class of arithmetical neo-logicist facts (facts grounded in some abstraction principle plus logic) and so on. Let $< Ar >$ be the set of all arithmetical truths. So we have:

$$[X*] \models < Ar >$$
$$[Y*] \models < Ar >$$
$$[Z*] \models < Ar >$$

It means that the different systems of arithmetical facts are arithmetically indistinguishable. Now, if we accept this, it is arguable that we are in presence of another fact:

$$[Ar-Ind([X*],[Y*],[Z*],\ldots)]$$

Namely, the fact that the arithmetical facts grounded on possible structure, proofs, abstraction principles, and so on, are identical from the arithmetical point of view. This complex fact has other facts as components (and a complex relation). Now, this fact should be in turn grounded since its components are not primitive facts; but what is, now, the structure of grounding relations? It should be something as the following:

$$\begin{array}{l}[\ldots structures\ldots] \triangleright [X*] \\ [\ldots models\ldots] \triangleright [Y*] \quad \triangleright [Ar-Ind([X*],[Y*],[Z*],\ldots)] \\ [\ldots proofs\ldots] \triangleright [Z*]\end{array}$$

Let us notice that in this case we have no phenomena of over-determination of grounding: every fact about structures, proofs, concepts, respectively grounds different arithmetical facts. These facts, *together*, ground the complex fact of arithmetical not-distinguishability. By transitivity of grounding, we can say that the fact of arithmetical indistinguishability is grounded by facts about structures, proofs, concepts, and so on. So, we get the first conclusion: we save Rosen's intuition about the grounding of arithmetical facts. And, our guide for finding the grounding relations is deeply connected with the concept of reduction of arithmetical truths. But we are not forced to choose just one true reduction nor to accept embarrassing cases of over-determination of grounding. There are many arithmetical* facts, all arithmetically equivalent.

We are now in the position to state our conjecture: from the grounding analysis of arithmetic we can say that the subject of arithmetic is constituted by the common features of different facts. It is easy to notice a structuralist flavour in this characterization; as Benacerraf puts, the subject of arithmetic is the common structure shared by all the omega-sequences. However, in case we considered the fact about the arithmetical indistinguishability concerns facts about structures too. And one could consider it a sort of meta-structuralism.

Perhaps there is room for a further conjecture. The starting point is the following reflection: [… Structures …], [… Proofs …],[…Concepts…] are extremely heterogeneous facts. How is it possible that they ground arithmetical facts which exhibit structurally identical features? Here, I propose a parallel which could help

us to refine our intuitions. Let us consider the well-known debate on 'natural' and 'artificial' consciousness:

[…Neurones…] ▷ […Cerebral consciousness…]
[…Microchips…] ▷ […Artificial consciousness…]

So, let us assume that these relations of grounding hold: the human consciousness is grounded in some (very complex) facts about neurons while an (hypothetical) artificial consciousness is grounded in some (very complex) facts about microchips. Well, let us assume that the two kinds of consciousness are, as a matter of fact, totally indistinguishable: we are in a context similar to science-fiction AI. We then say that:

[…Neurones…] ▷ […Cerebral consciousness…]

▽
[…not-distinguishability of consciousness...]
△

[…Microchips…] ▷ […Artificial consciousness…]

I think that this case is analogous to what presented in case of arithmetic. But here (maybe in a stronger manner than in the previous one) we have the strong feeling that the fact according to which speaking and interacting with a robot or with a human is the same thing cannot be a *brute contingence*. Why the artificial and natural consciousness are so similar, in fact, they are – some certain respects – the same thing? My conjecture is that there is some deep metaphysical fact which grounds the not distinguishability of the consciousness.

So, returning to our subject matter, there is some deep metaphysical fact which grounds the fact that facts so different show common arithmetical features. This fact is not the grounding of the fundamental facts about structures, proofs, or concepts. That is, it is not a supergrounding basilar fact. The reason is that the facts in question are too heterogeneous to have a common ground.

Of course, it is not easy to understand the nature of this fact; I can presume that it has to do with the concept of discrete and ordinate succession. For this, I will call it $[\Omega]$. Therefore, the final schema of grounding should appear as the following:

$[\ldots structures \ldots [\Omega]]$	$\triangleright$	$[X*]$	
$[\ldots models \ldots [\Omega]]$	$\triangleright$	$[Y*]$	$\triangleright [Ar - Ind([X*], [Y*], [Z*], \ldots)]$
$[\ldots proofs \ldots [\Omega]]$	$\triangleright$	$[Z*]$	

[Ω] is not a *separate* fact. If it were, we should admit that there exists, after all, a fundamental arithmetical fact, against the general reductive assumption made by Rosen. Of course, there is no problem, in principle, with this solution. But we think that [Ω] is a component of very different facts concerning very different kinds of entity. [Ω], so to speak, is the (arithmetical) *form* of all these facts (about possible structures, proofs, and so on) and it is this form which grounds the fact that, from an arithmetical point of view, all the particular arithmetical facts are not distinguishable. One can object that the existence of arithmetical forms of this kind seems precisely to be the sort of non-reducible fact about arithmetic. [Ω] has the explicative function, that is, of providing a good reason of why so different and heterogeneous facts share a common "arithmetic" feature. So, if we do not want concede that this is a brute contingency we have to postulate a reason for that. But we are not committed to the existence of a separate proto-arithmetic fact [Ω]. [Ω] seems to be the mode in which are "made" facts about structures or concepts of proofs. To emphasize that mode means to focus on their common features, namely, on their arithmetical properties.

6 Conclusion

In this paper we tried to extend the metaphysical approach of grounding to the philosophy of mathematics. As working hypothesis, we focused on the fundamentality (if any) of arithmetical facts. A provisional conclusion seems to be the following: both we consider the arithmetical facts as fundamental and we consider them as grounded, it seems unavoidable the reference to a common structure of all systems which exhibit arithmetical features. From this point of view, one can claim that, by focusing on the grounding relations between arithmetical facts, it is vindicated a form of meta-structuralism.

Lorenzo Fossati

Risk *vs* Logic. Karl Barth and Heinrich Scholz on Faith and Reason

Abstract: The paper analyzes a few questions of Heinrich Scholz on Karl Barth's dialectic theology: is it possible to view theology as a science? What is the meaning of a «theological proposition»? Which minimal formal constraints of meaningfulness shared with other sciences shall be observed in theological research? These issues are the necessary prerequisites for any rational discourse, hence for questions related to faith.

The paper aims at recalling the «classical» question of the scientific status of theology. I will start by illustrating how theology has been shaped in the previous century, taking the cue from Karl Barth and considering the origins and premises of his position; then I will try to outline some possible connections with other philosophical and theological conceptions (be them both in favor of or against) and eventually I will focus on the debate he sparked off between Barth and Heinrich Scholz, a less known and influential author, but nonetheless a very important one in the history of epistemology and analytical philosophy, at least (but not only) in the German-speaking field of research. The background of this investigation is the general issue of how and whether it is possible to deal with a theological question from a philosophical point of view. It is highly likely to argue the assumptions I will provide, but maybe this is the right way for it to be.[1]

1 The starting point

It is not controversial that the theology of the twentieth century springs from the crisis of the protestant *liberal theology*, whose history «ends» with Harnack and

1 One of the masters I am intellectually in debt is Sergio Galvan who taught me to use this kind of sensibility in research. I express my deep gratitude and esteem to him and I cherish fond memories of our discussions, including his annoyance at the most esoteric pages of some contemporary theologians or philosophers: in complete frustration we often asked ourselves «What does that mean?», echoing the most venerable «Was meinst du eigentlich?».

Lorenzo Fossati: Catholic University of Milan

DOI: 10.1515/9783110529494-009

Troeltsch: while looking for the *kerygma*, the *essential core* of Christianity, the historic-critical method of the liberal theologians, in its various manifestations, leads to an almost total destruction of the specificity of Christianity, i.e. its own object of inquiry. Once deprived of any historical stratification and theoretical Hellenization covering, such essential core should be the message of the Gospel: «God and soul, the soul and its God» (Harnack (1900), p. 68); the Gospel therefore does not announce the coming of the Son, but rather that of the kingdom of God. In the attempt to find what is specific in the Christianity, the liberal theology that makes use of other sciences' method seems:

(a) First of all to make it hard to distinguish the *Christian message* from a too secularized ethical imperative, no matter how elevated it is (Troeltsch (1998) p. 146: Christianity has the highest relevance for us as the «strongest and most concentrated revelation of personalistic religious apprehension»);
(b) Secondly it ceases to have its own *disciplinary specificity*.

There is then a double reaction from *dialectic theology* in its «restoring» intent:

(a) Primarily it has to be restored and guaranteed the characteristic of *Christianity* of being irreducible, compared to the other ethical proposals and religions: placing it at the first place in a «classification» that takes into account only gradual and not qualitative differences cannot be enough;
(b) Secondly, *theology* should stand out from other sciences and human discourses and set its own criteria of rigor and scientific validity, starting from its own object: its method cannot be borrowed from other fields of knowledge.

From this double point of view, therefore, we can probably better get the meaning of the «ganz Anderes», the «wholly Other» which Karl Barth stresses; from then on Barth's stance has become a miles stone for the theological thinking. Quoting Hobsbawn's expression, also the theological 20th century is short, starting in the 1922 with the second edition of *Römerbrief* (Barth (2010)); on the other hand the influence of this book is not over yet, so we cannot talk about short century at least for the following five years.

In Barth's opinion the Word of God is the object of theology. Given the two above mentioned aspects we can conclude that:

(a) As far as the *specificity of Christian religion* is concerned, compared to other religions and moral systems, the judgment of God is the «no» to the world and to human history, marked with sin and death, and it is (dialectically) overcome by the «yes» of the Incarnation and Resurrection of Christ: for the world, His Word is «crisis». Since it is not possible a way that begins with the man and leads to God (on the contrary, it is only the Grace coming from Him

that allows the human to establish a contact with Him), it is impossible to include the Christianity in the human attempts to reach the absolute.

(b) As for the *specificity of theology* as a science, faith is interpreted in Kierkegaard terms as a leap, a risk, a void space for Grace: the discourse of theology, therefore, should stick to its object (that is the Word of God, definitely spoken by Christ) and the criterion of its scientific validity has to be identified in the «objectivity» [*Sachlichkeit*], in its appropriateness to its object that has to be deciphered, starting from the text, the Scripture.

Barth of course went further this dialectic stage and expressed his new ideas in *Kirchliche Dogmatik*, his unfinished *magnum opus*, on which he has worked from 1932 to death, in 1968. The work contains relevant differences compared to *Römerbrief*, particularly the «otherness» between God and man is moderated. However, as for the two questions we brought up, Barth's position gets radical and does not tone down:

(a) On the one hand, the theology of the Word of God becomes rigorous, as a Christology that states «the recognition of the humanity of God on the basis of the recognition of His divinity»: God always remains «Other» than the world, but He gets close to it in terms of alliance and reconciliation thanks to His Word in Christ. This means that a connection between *Christianity* and other – *Menschliches, Allzumenschliches* – approaches is still impossible.

(b) On the other, *theology* is nonetheless linked to the community of the Church and «only in this way it becomes a possible science and gets its meaning», proving itself to be a science neither independent nor without presuppositions.

At a closer look then the fundamental difference from the first assumption, i.e. the convergence between God and man, does not lie in a mitigation of the original thesis (which is anyway present), but rather in a radicalization and a deeper consequentiality with respect to the theological object, i.e. the nature of the Word, in Christ addressed to the man.

2 Dualisms

Those philosophers who examine Barth's position can find in it many typical themes of the contemporary thinking, even though it is better to avoid a too simplistic overlapping: Barth indeed does not want to be a philosopher (see Barth (1960)) and criticizes the theology for being too close to the philosophy of religion

or culture (which is the mistake of liberal theology). Nonetheless there are some theoretic premises that the philosopher can consider as being part of his field of study (as Pannenberg (1973) did, but in the opposite direction, being a theologian who deals with philosophical issues).

As paradoxically as it may sound I would like to introduce a parallelism; Barth starts by doubting the historicist results and approaches to Christianity and theology; but when he focuses on the irreducibility of the Word of God and puts the theology against any other science in a radical dualism, it echoes the epistemic dualism, developed within the German historicism, between explanation [*Erklären*] and understanding [*Verstehen*] (see Dilthey (1883)).

While the «sciences of nature» [*Naturwissenschaften*] try to «explain» the physical world by connecting each single fact to the generality of laws and putting them into a causal chain, the «human sciences» [*Geisteswissenschaften*] have to «understanding» a human and not physical phenomenon, that it is then not a simple «fact», but an «event». In the humanities – history first of all, then psychology and philosophy – man is at the same time subject and object of the analysis and since he cannot abstract *himself*, he is not able to follow an empirical method of research.

Such *epistemic dualism* is clearly based on an *ontological dualism* – between subject and object, man and world – a master of whom we can identify at the origin of the so-called «modernity» in Descartes; this of course holds only for a rough connection, since the references to Plato or to the Aristotelian physics, claiming that a celestial body is an autonomous object of research compared to the sublunary one, could be equally valid.

Here then it is possible to find an interesting hint, if we consider that one of the crucial turning points of the 17th century «scientific revolution» lies exactly in the refusal to treat in a different way the two kind of movements (even though it is now acknowledged that even in the Middle Age some authors had followed the same direction). A unified theory on the physical world, able to follow a unified method, makes it necessary the affirmation of an *epistemic and ontological monism*. I believe that the Diltheyan dualism is easier to understand if we do not interpret it as the attempt of an unlikely pre-modern recovery, but as the attempt to label as science an area that seems to be cut off from the criteria of scientific rigor and accuracy or, at least, extremely downsized from a reductionist standpoint.

However these methodological distinctions, that wanted to extend and not to limit the scientific enterprise, are historically at the basis of the hermeneutical thinking that criticizes the empirical science in general terms: it does not only *distinguish* between explanation and understanding, but it *subordinates* the first to the second, considering explanation a derived – or degenerated – kind of

understanding. Let us consider the topic of «technology», the pure *rechnendes Denken* that remains at the level of things and that Heidegger opposes to the *Andenken* looking for the meaning of Being; moreover let us think about the result (maybe less radical and explicit) generated by the critics of the *Frankfurter Schule* to the *instrumentelle Vernunft* of the Enlightenment or the approach to the problem in the fundamental work for the contemporary hermeneutics, *Wahrheit und Methode* by Hans Georg Gadamer.

Therefore I think it is useful to compare Barth and Dilthey, at least in order to understand the nature of their likewise «theoretical maneuver», since the two of them share the need of a distinction and an irreducible autonomy to both the object of their discipline and the method that characterizes such discipline: here emerges the typical Kantian issue on the conditions of possibility of a (precise form of) knowledge, hence the goal is not *stricto sensu* «anti-scientific», but on the contrary aims at extending the field of the science, and not at reducing it.

And what about the possible convergence between Barth and hermeneutics in terms of «subordination» – in one case of the explaining to the understanding, in the other of the reason to the faith? I will try to illustrate it in the following paragraph.

3 Theological legacies

If establishing a distinction between different spheres seems to be a defense *of* the science, rather than *from* it, is nonetheless true that it is legitimate to ask whether this idea is, in Karl Popper's words, a *strategy of immunization*: a theory cannot be put under criticism from «outside», because such criticism is rejected and considered non pertinent… only because it could be fatal! Thus the argumentation comes full circle because the claim that the criteria of scientific validation of empirical sciences do not hold for human sciences would derive not only from the *ontological and gnoseological dualism* we mentioned before, but primarily by a Christian *theological legacy* that places the man at the centre of the world. Karl Löwith (1949) comes to similar conclusions, when he affirms that the philosophy of history is nothing else that the secularized version of eschatology, whereas Ernst Topitsch (1958, 1990) goes further and does not connect the spirit-nature dualism just to the theological or religious legacy of Christian religion, but even to the mystery-magic legacy of Gnosticism, to the ecstatic-cathartic representations of the mythical thinking and of the ancestral shamanism, of which both metaphysics and Christianity would be epiphenomena.

Whatever it may be true, anthropocentrism, far from being the conclusion of an ontological or metaphysical argumentation, would then be a missing

assumption of the investigation, which would eventually compromise it. But here we have to cope with the recurrent problem that shows up every time that a reductionist request is brought on the table: those who support it deny that there are differences that exclude it and blame the opponent for mystification or dogmatism, while the opponent in his turn objects that it is the reductionist the blind and simplistic between the two of them.

If this can hold for the methodological question (we can quote *The Poverty of Historicism* by Karl R. Popper (1957), or the so-called *Positivismusstreit* of the Sixties between Frankfurter Schule and critical rationalism, see Adorno and et al. (1969)), as regards theology and theology as science, the issue crosses the need of a lack of presuppositions [*Voraussetzungslosigkeit*] and of a lack of value judgments that for Weber are required to the scientific enterprise (see Weber (1919)). From this point of view it is clear that theology is exposed to a checkmate, since it cannot meet none of the two requirements, thus placing itself ipso facto outside the domain of science.

Popper [1963a] criticized the «manifestation theory of knowledge» that aims at combining classical rationalism and empiricism in order to look for a foundation of the knowledge, substituting the religious authority with the authority of evidence, be it drawn from the reason or the empirical data. When Hans Albert (1968) translated this concept he used the expression «*Offenbarungsmodell der Erkenntnis*», thus making immediately evident the association (due basically to the terminological identity) with the religious revelation, that in this way becomes the archetype of the «dogmatic» thinking, i.e. an alternative to the fallibilistic and critical approach that instead proceeds by means of conjectures and refutations, and that includes the scientific enterprise as its most mature accomplishment. It's William Warren Bartley III (1984) that claims the contraposition between the critical rationalism and the «retreat to commitment»: critical rationalism gives up committing to a justification or to a final grounding of the proposed theories because the result of the task would be necessarily judged as dogmatic and therefore irrational. Instead Popper's methodology identifies the scientific character in the falsifiability and, by extension, the rationality in the possibility to be full-blooded criticized (including critical rationalism itself).

In this sense Barth becomes the privileged interlocutor for Bartley: the theologian assumes the impossibility of «value-freedom» and the lack of presuppositions in science, thus placing at the same level any human cognitive enterprise and rejecting the accusation of dogmatism for the theology, because any science can be charged by the same accusation. If it crossed someone's mind to reject the idea by opposing an value-free and «scientific» (!) conception of science, Barth could just reply: «tu quoque» – all dogmatic, no dogmatic.

For the critical rationalism then the only possible way is to renounce to an impossible need (the grounding) in favor of a practicable one (the critique); Barth's strategy instead goes on the opposite direction, towards a radical grounding, starting from the irreducibility to its own object (the Word of God) and being consequent and adherent to it. The criterion of the scientific status of theology has to be identified from the reality or actuality of the subject-matter [*Sachlichkeit*] or the adequacy of the discipline to its subject-matter [*Sachhaltigkeit*, *Gegenstandsgemäßeit*] (see Barth (1932–1967): I/1, 7).

As I said before this conclusion may seem, from a «rationalistic» point of view, quite similar to the «disqualifying» result the hermeneutics comes to when it deals with the explaining issue: there explanation is subordinated to comprehension, here reason is subordinated to faith.

4 Theology as a science

In this respect, if we want to outline a general framework of the positions we are analyzing (provided that we cannot show all their pluralistic richness), there are two basic models in the development of the self-comprehension of theology as a science:

(a) First of all, the model that considers it a *theoretical science*, a speculative science on God: *sacra doctrina* has to come from axioms taken from the articles of faith, the dogmas. Among them, with reference to *Analytica posteriora* (I, 9, 76a), for Thomas Aquinas theology is a «derived science» that takes its own principles from a «science», i.e. the evident knowledge of the articles of faith that God and the angels possess, which is assumed by theology through *lumen fidei* and not thanks to *lumen rationis*.

(b) The second model traces back to Duns Scotus and considers theology as a *practical science*: this model aims at restoring one of the first self-comprehension of theology as *sapientia* (already present in Augustine) that combines in itself both the theoretical and practical aspects, treating the centrality of God as the *Summum Bonum* orienting the human affection and action, rather than a simple object of knowledge. *Cognitio practica*, in Scotus' system, is superior to any other speculative knowledge, since it is about the knowledge of an end.

In the Scholastics we can find this alternative, that anyway – I would suggest – obtains the right relevance only later on, if we consider it as the source of the so-called «anthropologic turn» in theology of the 20th century: if in

the context of Scholastics God was affirmed with reference to a theoretical (cosmological-metaphysical) demonstration, in the new understanding the object of theology is not only *God*, but the *relationship between man and God*. Therefore we can see in the «anthropologic turn» not only a simple fallback or a strategy of immunization, but the recovery of a different comprehension of theology as a practical, non theoretical science (in the established terms). Such turn is not dictated from the outside, but it grows up autonomously in the theological grounding of the ontological and teleological structure of the human being in relation to God.

On the subject of the scientific status of theology another relevant contraposition has become central in the context of Enlightenment:
(a) on the one hand we have *natural* theology and religions,
(b) on the other, *positive* (or *revealed*) theology and religions,

based on the model of the relation between natural and positive law. Lessing (2001, §§ 9, 11) concludes that «all the positive and revealed religions are consequently equally true and false» and that «the best positive or revealed religion is the one containing few conventional additions to the natural religion and limiting as little as possible the good effects of the natural religion». As regards such conclusion one can object that the evaluation could be opposite, if we consider the concept of *natural* religion as a mere abstraction that comes from the *positive* religion which is not, then, arbitrary but the only real and determined. The same problem arises in the tension between the *natural-metaphysical* (rational) theology and the *positive-revealed* (dogmatic) one: is it really possible a natural-metaphysical theology without presuming somehow a dimension of faith where to actually find the believer? The question triggers off the alternative between theoretical and practical theology with reference to the question whether it is possible a theological science, regardless of the subject that practices it and regardless of the Church.

Barth answers with a clear and double «no»: there is no theology that is not revealed, there is no theology that is speculative. Taking the faith of the Church as the foundation of theology seems however a clear example of «immunization of the critics», since only believers could practice such discipline, because they accept certain presuppositions that the rational approach cannot investigate and they start from their own self-comprehension based on those very premises. To those who object to this auto-referential «seclusion» Barth would reply that even the opponent would end up being in such context, since *fideism* is the necessary situation for any kind of investigation, including the *rational*, *scientific* and *critical* one. We also recall the widely debated expression by Karl Popper (1945) in *The*

Open Society and its Enemies which places at the origin of the critical rationalism an «irrational faith in the reason».

Whether it is «ecclesiastic» or not, theology would not exist without premises, since no human cognitive enterprise is able to satisfy such need. I believe that this is the point where the ideal connection between hermeneutics (in a wide sense) and theology (in a modern «20th» sense) takes place.

5 Scholz's requirements

Barth, as we have seen, is ready to include theology in the realm of sciences, at least in the *Kirchliche Dogmatik* period, and this is made possible because the scientific status is defined by two aspects: the adequacy of the discipline to its object and the rejection of irrationality, an accusation that positive or «natural» sciences can no longer address to theology nor to any other form of knowledge, that is, to no one. It is useful to remind the more epistemological work by Barth, *Fides quaerens Intellectum* [1931], where in the comparison with Anselm the Augustinian topic of «*credo ut intelligam*» is restored; in this principle the priority is given to the *credere*, but a role is assigned also to the *intelligere*. Including theology into the sciences takes nonetheless its high toll and for many aspects it is unacceptable for Barth's opponent, who is supposed to criticize in the first place Barth's criterion of scientific validity.

This is the starting point of the objections that Heinrich Scholz makes in his essay *Wie ist eine evangelische Theologie als Wissenschaft möglich?* [«How is an evangelical theology possible as science»], appeared in 1931 on the journal *Zwischen den Zeiten*, the «official» organ of dialectical theology. The adequacy to the object [*Sachlichkeit*] in fact is not a question that we can figure out

(a) independently from a *formal* criterion that establishes what and how can be defined as the object of a science;
(b) independently from the need of *controllability* of its propositions.

Such requirements make it possible a judgment on the objectivity and consequently on the truth of a proposition, since, if they do not exist, a «dogmatic» – non scientific – result is unavoidable: any discourse could claim to be scientific, just in so far as it proves to be «faithful» to its own object which, however, is part of the discourse itself: astronomy is thus equal to astrology.

The minimal and uncontested requirements [*nichtumstrittene Mindestforderungen*] that, according to Scholz, can constitute a better criterion for labeling any discipline as a science have a pure *formal* character, that is, they

leave out of consideration the specificity of the object, of the *matter* of a specific science:

(1) The *propositional* postulate [*Satzpostulat*]: «Apart from questions and definitions, in a science only propositions can appear, that is assertions that are said to be true» (Scholz (1931), p. 19); in this postulate it is clearly implied the principle of non contradiction, which de facto is the minimal condition of possibility of a science, insofar as it seeks the true and avoids the false.
(2) The *coherence* postulate [*Koherenzpostulat*]: «We can speak of science only if all the propositions that are part of a science can be formulated as propositions of the objects of a certain specific domain» (*ibi*: p. 20); there exist then a unified domain of objects that excludes all the non-pertinent propositions.
(3) The *controllability* postulate [*Kontrollierbarkeitspostulat*]: the fact that the affirmations of science have a demand of truth (according to 1) is not enough; instead «we have to request such demand to be controlled in some manner» (*ibi*: p. 21). Controllability however has to be interpreted in a more general sense than the one proposed by the Neo-empiristic principle of verifiability, which obviously cannot be applied to theology (and maybe to anything else – a crucial point Scholz was well aware of in 1931).

Scholz adds also two other «contestable» [*umstrittene*] requirements:

(4) The *independence* postulate [*Unabhängigkeitspostulat*] refers to the absence of presuppositions in the propositions of a science. According to it «for the propositions of a science it is not acceptable that they are said to come from the pressure of some prejudice» (*ibidem*). Here Scholz goes back to Aristotle in his *Analytica posteriora* (I) and does not claim that science could assume anything; this would be absurd even for mathematics, that on the contrary formulates its own requirements in such clear and precise [*klar und pünktlich*] axioms that they make it the model of science; we are here talking about the necessity that the judgment is based on an objective demonstration and not on a subjective and uncontrolled prejudice.
(5) The *concordance* postulate [*Konkordanzpostulat*], according to which only propositions that do not contradict the true propositions of other disciplines can be part of a science; particularly, Scholz refers to «'our' physics and 'our' biology» (*ibi*: p. 23); the domain of knowledge has to be *unified*, as well as the world that is the object of such knowledge, and both have to respect the principle of non contradiction.

The two last postulates proposed by Scholz have a weaker grip in theology since the discipline is in a particular critical situation. For example, regarding (5)

the very concept of *miracle* contradicts by definition with what the empirical science admits; for the (4) the personal adhesion to the contents of faith is not only identifiable with a *presupposition*, which is structurally necessary but at least objective, but with a subjective *prejudice* that determines and grounds the demand for truth of the believed propositions.

Turning to postulate (4) and Aristotle Scholz mentions also the maximal requirement that we can demand to science, based on the model of mathematics:

(6) The *axiomatic* character, on whose basis «propositions of science should be able to split into two classes of propositions: the propositions where *being true* is a presupposition belong to the first class; we call them the *principles* [*Grundsätze*] or *axioms* [*Axiome*]. The propositions where being true is *deducted* [*deduziert*] or *proved* [*beweisen*] on the basis of the *being true* presupposed by axioms belong to second class. We call these propositions *theorems* [*Lehrsätze oder Theoreme*]» (*ibi*: p. 24). Dogmatic theology takes hits propositions from the creed of the Church, and therefore could in a certain way follow such postulate, but still the second aspect regarding the demonstration of theorems, remains problematic: «Affirming is easy, demonstrating is difficult, so difficult that today many philosophers do not know any more how it is difficult. And how it is beautiful, *because* it is so difficult» (*ibi*: p. 29).

Evaluating the minimal conditions and their implications, as well as the difficulties in applying them to theology, Scholz's conclusion appears to be non-positive: how is it possible to control the truth of the fundamental theological propositions, i.e. regarding God and Christ, the world and the man, as they appear in the Gospel… without taking into account the Gospel itself? For Scholz it is indeed impossible and «the only thing we can try to do is to seek the foundation somehow able to support [*stützen*] the *faith* in the truth of this propositions» (*ibi*: p. 48): the domain of science then seems to exclude an Evangelic dogmatic which can only exist as «a personal profession of faith, in its strictest meaning, excluded by any mundane evidence» (*ibidem*).

6 Barth's risk

A year later, in the first volume of *Kirchliche Dogmatik* [1932–1967, pp. 6ff], Barth categorically rejects the conditions expressed by Scholz: «We cannot grant anything without betraying theology, because any concession would mean abandoning the theme of theology». He does not exclude only the coherence and

controllability postulates, but maintains that the principle of non contradiction is only partially admissible.

Barth, however, does not want to share Scholz's conclusion, according to which it is unlikely for theology to be a science, since it is more similar to a subjective profession of faith, independent from any mundane and objective evidence. Barth, on the contrary, claims for theology the scientific status since

(a) it is a human strive towards truth,
(b) it pursuits a «specific object»,
(c) it follows «a way of knowing consequent per se» (*ibidem*).

But what is the meaning of the last statement (c) if we do not admit the principle of non contradiction (see 1 in Scholz)? And how can we expect to «account for everyone» without accepting the controllability postulate (see 3)?

Such questions seems to confirm Scholz's doubts about a theology «that interprets faith as a risk in such a radical way that one cannot predict anymore how it is possible to reach any propositions starting from this risk [...] that can be said to be true» (1931, p. 39). It does not make sense to ask whether a risk is true or false... In one word, it is a *reductio in mysterium*.

Barth too in 1930 referred to risk when he defined theology as «the science of faith», being «free obedience» that represents «the risk of a highly uncertain obedience» (p. 384); the theologian does not possess any evidence able to prove the consistency of his enterprise, «but he feels the Word of God and meditates on it» (*ibi*: p. 383). The point is philosophically crucial, since the possible scenario is problematic.

First of all, if the fundament of theology is the risk that the theologian takes on as a (human) subject, then the absolute priority of the object, i.e. God and His Word, is brought into question, but this is the very opposite of the purpose from Barth's perspective.

Secondly if the starting point of theology is the positivity of Revelation, but theology comes to the Word of God, taking its cue from the risk of faith the theological consciousness takes on, then an irrational subjectivism lacking any justifications and giving up any intersubjective and rational discourse takes place – a intersubjective and rational discourse considered mundane and thus able to corrupt its own «radically other» object. In short, did not we end up parceling out the object into its infinite risks taken on by the believers towards the object, «risking» (!) dissolving it?

7 Second round

So far we can conclude that both for Scholz and Barth it is not possible to label theology as science in its «ordinary» meaning, even though (and it is relevant) in Barth's opinion, theology is still a science and not a simple profession of faith, since he reshapes the concept of science so that he can include the theology in it (or maybe the other way round, since he defines such concept, starting from theology, as Bartley maintains).

On the one hand, then, Scholz and Barth agree that theology has nothing to do with science, on the other, however, they do not agree on the general definition of «science» and on the possibility to give theology a scientific status; moreover Barth specifies that theology has nothing to do with the *other* sciences, even though it is scientific itself: the first part of his conclusion is similar but the second part is opposite, since it comes from other premises: this is a very important detail since what matters here is how one maintain a thesis, rather that *the fact* that it is maintained.

In 1936 Scholz turns back to the question with *Was ist unter einer theologischen Aussage zu verstehen?* [«What is to intend by theological proposition?»], i.e. his contribution to the collection of papers for the 50th anniversary of Barth, where his opposition is neater, or better, it is expressed in a more precise manner.

The starting point for Scholz (1936, p. 25) is the following: given that it is generally demanded for an proposition to be able to be true or false, in a theological proposition it is the result of a determination [*Festsetzung*] (*ibi*: p. 27) and an agreement [*Verabredung*] (*ibi*: p. 28).

On the contrary, for Barth a proposition is theological:

(a) if its object is God or it is referable to a proposition on God,
(b) but first and foremost if «this proposition is not a rational proposition» (*ibi*: p. 33).

This last characteristic deserves some insights, because of its apparent paradoxical nature, and Scholz claims that he can conclude that for Barth a proposition is «rational»

(c) «if it appears evident to the 'natural' man» (*ibi*: p. 34).

Since we have to establish what is the «natural man» in Barth's terms, we have to infer that

(d) «with natural man we mean a man to whom a theological proposition in Barth's sense does not appear evident» (*ibidem*).

Scholz's conclusion is in between ironic and desolate: «It is impossible to become more circular than that». What leads to such circle is the Barthian refusal of the proposed minimal requirements and his claim that the theological proposition is a proposition on God but non rational. Furthermore, another «danger» is the claim to explain the non rationality of theology by turning down the deductive reasoning, since «nobody will forbid a theological proposition from having some logical consequence [*logischen Folgen*]» (*ibi*: p. 36), at least because «any logic consequence of a theological proposition will be a theological proposition itself» (*ibidem*). The «risk» Barth talks about, i.e. the opposition to rationality and logic (that is the expression of rationality par excellence), does not lead only to a circular and paradoxical position, but turns into a «danger», because in Scholz's opinion «it is, in any case, dangerous to deny a thinking being the deduction [*Schließen*]» (*ibidem*). This would mean to deny «the most essential instrument we have as thinking beings to make clear for ourselves [*um uns klar zu werden*] what we think. […] But if we were created by God, then we have to conclude [*schließen*] that He wanted us to make use of the intellect He gave us, so that we did not appear bad or unfaithful holders» (*ibi*: p. 37).

To sum up, Scholz rejects the idea that it is enough for a proposition to be controllable or demonstrable in order to be considered as non theological: he insists that it is impossible to renounce to logic, thus turning again to the epistemic minimal requirements proposed before; at a deeper look they just make explicit what is implicit in the very concept of assertion and proposition, that is in the logic of propositions:

(1) every proposition always affirms something as true and consequently it excludes the contradiction and non truth (according to the *propositional* postulate);
(2) every proposition always refers to something else that is always distinct from it; different propositions can then refer to the same object and can be considered as a description of it, given that they respect the non contradiction that makes them co-possible (according to the *coherence* postulate);
(3) every proposition reveals itself as an hypothesis on an object that can correspond to it or not, make it true or false (according to the *controllability* postulate).

From the very notion of proposition we infer that a proposition is meaningful only if it shows somehow the condition under which it is true: this however does not mean for Scholz a servile acceptance of verificationism, since this requirement does not admit only an empirical controllability; Scholz shows in fact the need of a kind of provability: even the theological propositions cannot avoid logic. If Scholz maintains that «only through deduction we are generally led from obscurity to

clearness [*aus dem Dunklen ins Helle*]» (*ibidem*) does that mean that he rejects any acceptation of theology? I would not say that. What it is left out is just a theology that is *exclusively* positive or revealed.

We can hear in this regard an echo of Thomas's saying *gratia non tollit natura sed perficit*, which coincides with the condition of possibility of any human approach to God, be it due to human action or to God's Grace. Every science, roughly speaking, being a human enterprise, has to be carried on in a human way, that is with logic, at least in as much as it claims to have a cognitive character: the scientific status of theology is to be found – or not found – on this point.

8 «Aus dem Dunklen ins Helle»

In conclusion I think that the analysis of Scholz's objections and the history of this debate does not have just an «archeological» meaning. These objections may seem very *naïf*, and generally speaking a theologian can always consider the observation made by a philosopher or a logician reductive or non pertinent. I do not think so; first of all theology, as any other human cognitive enterprise, can take advantage of objections, but above all because our researches have always aimed at passing «from obscurity to clearness», as Scholz said. We have to admit tough that not every philosopher believes such ideal of clearness to be fundamental (there is a great deal of examples of that), therefore analyzing this debate could be strictly useful from a philosophical and not only historical point of view, as a sort of case study of the contrast between analytical philosophy and hermeneutics.

We would not discuss any theoretical problem if we did not believe that it is important to look for the conditions of possibility of a «rational» solution (I would rather say «scientific»), and this should also apply to theological problems. In the Gospel, after all, it is said: «*Sit autem sermo vester: "Est, est", "Non, non"; quod autem his abundantius est, a Malo est*» (Mt 5;37).

Aldo Frigerio

On the Ontology of Biological Species

Abstract: In this paper, two different ontological views concerning biological species are analyzed. On the first view, species are universals instantiated by the members of the species, while, according to the second view, species are complex individuals formed by the members of the species. An alleged decisive argument in favour of the second view is based on the fact that biological species evolve, while abstract entities such as universals cannot change. It is shown that this argument is far from decisive because other kinds of abstract entities such languages are said to evolve. It is, then, illustrated in which sense we can say that abstract entities, such as biological species and languages, change.

For a long time, biological species were considered universals or kinds of which the organisms that are members of the species are instances. Indeed, the relationship between species and organisms belonging to the species was regarded as the paradigm of the relations between universals and particulars. This view, which dominated from Aristotle to Linneaus, was undermined by Evolutionism. According to this theory, biological species have a temporal beginning and a temporal end, and they evolve and change. Abstract entities, such as universals, are usually conceived as timeless and, thus, as entities that have no beginning and no end and cannot undergo changes and transformations. Many scholars have concluded that biological species cannot be kinds and that every attempt to deny this was anti-Evolutionist and fixist because it would have implied the negation of the thesis that species evolve[1]. A new ontological interpretation of species and of the relations between species and members of species was necessary. Michael Ghiselin and David Hull, among others, have claimed that species are complex individuals and that organisms belonging to a species are parts of that species[2]. Other examples of complex individuals to which species can be compared are firms. Of course, individuals have a beginning and an end, and they can change over time. Therefore, they seem to be the right ontological category into which to place biological species.

1 For such a view, see Hull (1965a), Hull (1965b), and Sober (1980).

2 See, in particular, Ghiselin (1966), Ghiselin (1974), Hull (1976), Hull (1978), and Ghiselin (1987).

Aldo Frigerio: Università Cattolica, Milan

DOI: 10.1515/9783110529494-010

The view that species are individuals suffers from problems that can hardly be overcome[3]. However, it is not my aim here to review these problems. This essay has another aim: the main reason for adopting the species-as-individuals view is that species evolve; therefore, we need to investigate whether the evolution of species is compatible with the thesis that species are kinds. We will see that we actually speak of the beginning, extinction, and evolution of entities different from biological species that are undoubtedly abstract. Therefore, the evolution of species can hardly undermine their abstract status. Moreover, it will be shown how timeless entities can have a temporal evolution. Timeless status and evolution do not seem to be, in the first analysis, compatible: our aim is to show that they actually are. Among the abstract entities usually seen as evolving, in light of the philosophical education of the author of this essay, I will pay particular attention to languages. Other examples, such as theories and cultures, would be equally suitable for the aims of this paper.

This essay is structured as follows. In Section 1, some definitions of the species-as-kinds view, which dominated from Plato to Linneaus, are analyzed. In Section 2, the concepts of species as individuals and the reasons they were adopted are examined. A decisive reason seems to be that species evolve, while kinds are timeless entities. In Section 3, it is shown that biological species and languages are similar in several aspects and that the fact that languages evolve does not prevent us from considering them abstract entities. Therefore, the argument based on the evolution of species in favor of the view that species are individuals is much less strong than appears at first glance. In Section 4, it is shown how it is possible that abstract entities, such as languages and biological species, can have a beginning, an end, and an evolution. Section 5 contains some concluding remarks.

1 Biological species as kinds

Traditionally, classification is regarded as an operation that assigns an individual to a class on the basis of its properties and features. The basic idea is that individuals have different properties and features, and they can be classified by virtue of the properties they exhibit. Kinds are defined as properties or as sets of properties and features that can be instantiated by individuals. It is well-known that different philosophical positions concerning the ontological

3 For a review of such problems, cf. Stamos (2003).

status of kinds are possible. Realist views conceive of kinds as entities that exist autonomously, while, according to conceptualist views, kinds exist solely in our minds. Intermediate positions are also possible. Here, no stance concerning the ontological status of kinds will be taken. It will suffice to say that, whatever their ontological status, they are usually considered abstract entities that are not existing in space and time, unlike the individuals that instantiate them. In many cases, the individual's membership to a kind is defined on the ground of a set of necessary and sufficient conditions that the individual must satisfy in order to belong to the kind. The individual belongs to the kind only if it satisfies these conditions. Such conditions are often identified with the possession of the set of properties and features that identify the kind. The possession of each of these properties is considered a necessary condition for belonging to the kind, while the possession of all the properties is thought to be a sufficient condition for the membership of the individual to the kind.

More flexible views of the membership to a kind are, however, possible. Rather than a question of yes or no, the membership to a kind can be regarded as a question of degree. In this case, to belong to a kind, the individual does not need to possess every property that forms the kind but only a substantial number of them. Here, I will not deal with the question of whether or not the membership to a kind is a matter of degree. For the aims of this essay, the important thing is that the membership to a kind is evaluated on the basis of a set of properties or features that the individual must possess.

A biological organism's membership of a species was interpreted along these lines until the advent of Evolutionism. Biological species were thought of as abstract kinds that were instantiated by biological organisms. This sort of approach goes back, at the least, to Plato. The Latin word species is a translation of the Greek word *eidos*, which means "idea" or "form". It has the same root as a Greek verb meaning "to see" and it is related to the idea of what is seen and of visible form. Plato used this word, which was very common in everyday language, to refer to a new metaphysical category – that is, what would be called *universalia ante res* in medieval times, to entities graspable only by the mind, which coincide with the abstract and immutable essences of perceivable things. Concrete and perceivable individuals are only more or less successful copies of these immutable Forms.

However, Plato did not apply his metaphysical view explicitly to biological species. Aristotle was the first philosopher who focused his attention on the biological realm. In *Posterior Analytics* and *Metaphysics*, Aristotle seems to endorse a classification method based on the division: to grasp the essence of a thing, one needs to understand its generic nature and the features that distinguish that thing from the other things that share its generic nature. These features are

the specific difference, and they coincide with the species to which the thing belongs. However, in his biological writing, such as *Parts of Animals*, Aristotle is far less clear regarding the method for classifying biological organisms. He acknowledges that it is often difficult or impossible to draw clear borderlines between the biological and nonbiological world and among the biological kinds (Aristotle 588b4–13) and that some organisms can belong to a kind for certain aspects and to another kind for other aspects (Aristotle 588b 13–17). Moreover, his main aim in biological writing is likely to investigate the principles at the basis of the distribution of characters or parts of animals. In particular, his project is to study the necessary correlation among organs and which organs are functionally necessary to the lives of certain kinds of organism[4]. Aristotle is committed to the study of the relations among the parts of an organism, of their function, and of the relations between organisms and the environments in which they live.

Some interpreters see tension between Aristotle's logical and metaphysical writings on the one hand and his biological writings on the other hand. There would be inconsistencies between the more aprioristic and essentialist approach of the first group of writings and the more empirical and pragmatic approach of the second group. Moreover, in Aristotle's writings, *eidos* would have two senses that are not completely compatible: the first sense is logical and conceptual and is very close to Plato's and Pythagoreans' sense, expressing the form of objects; the second sense is related to the dynamic and vital principle of the organization of biological life[5]. Other interpreters do not see such a contrast between the two groups of Aristotle's writings and believe that the view expressed in the biological writings is not in opposition to the view expressed in the logical and metaphysical writings[6].

Whatever the solution to this dispute, Aristotle's reception in the Middle Ages and in the ensuing centuries was often mediated by Neoplatonic readings, such as that of Porphyry's *Isagoge*. Commentaries were focused on his logic and metaphysical writings, and their aim was often to reconcile Aristotle with Plato. Even after the discovery of the whole Aristotelian corpus in the XII century, very little attention was devoted to the biological writings. By consequence, the method of the division for genus and specific difference was inherited by the Renaissance naturalists and was followed even after the scientific revolution.

4 For such an interpretation of Aristotle's biological writings, see Wilkins (2009) and Richards (2010).

5 On this contrast, see, in particular, Richards (2010).

6 Cf., for instance, Stamos (2003), pp. 107–111.

Studies focused on the biological world began with medieval bestiaries and herbariums and were carried on by the naturalists of the Renaissance period. One of the most important figures was Andrea Cesalpino (1519–1603). In his *De plantis libri XVI* and *Questionum Peripateticarum*, he refers explicitly to Aristotle by arguing that knowledge consists of the predication and hierarchy of universals. He is also Aristotelian in his method of classification, which is grounded on the two fundamental functions of plants – nutritive and reproductive. It is on the basis of the systems of organs and features, which have the nutritive and reproductive functions, that Cesalpino divides plants. For the first time, he proposes a system of cataloguing that is not based on the alphabetical order of the names of plants, as in the previous herbariums, but hierarchical, even though his hierarchy is limited to species and genera and does not include kingdoms, classes, and orders, as in Linneaus' system[7].

Cesalpino's work was carried on by John Ray (1627–1705) and, above all, by Linneaus (1707–1778). The latter employed sexual ("fructification") characters in plants for his botanical classifications, following the Aristotelian tradition, according to which functional characters are central to taxonomies. Linneaus' system is hierarchical and, therefore, follows Aristotle's method of division. In contrast to what is usually believed, the tradition I sketched out here is not fixist and anti-evolutionist in principle. For instance, Linneaus allows for some new biological species to arise through hybridization[8]. In his *Systema Naturae* (tenth edition, 1758), Linneaus writes that God created an original individual or mating pair for each genus and that new species were produced by intergeneric crosses. In the thirteenth edition of 1770, he goes even further and hypothesizes that the original breeding pairs or individuals might instead represent orders, rather than genera, and that even new genera, as well as species, might be formed through hybridization. However, Linneaus' evolutionism has a twofold limit. In the first place, he does not allow for methods of the formation of new species different from hybridization. The possibility of mutation in the transmission of hereditary characters had not yet been discovered, and it was thought that a couple of organisms belonging to a species could not generate organisms belonging to a different new species. In the second place, given the couples of individuals belonging to each genus, the species that can arise by hybridization from these couples are rigidly predetermined[9].

7 Some information on the work of Cesalpino can be found in Larson (1968) and Wilkins (2009), pp. 56–57.

8 Cf. Larson (1968), p. 293 and Eriksson (1983), pp. 94–95.

9 Cf. Ereshefsky (2003), p. 207.

Georges-Louis Buffon (1707–1778) was Linneaus' main opponent. As regards the classification of living beings, Buffon proposes a criterion that differs from that of Linneaus, who was mainly interested in the characters of individuals. He, instead, embraces the criterion of breeding with fertile progeny: an organism belongs to the species *S* if it can mate with organisms that are members of *S*, thereby producing a potential limitless progeny[10]. This criterion will be at the basis of the biological concept of species advanced by Theodosius Dobzhansky and Ernst Mayr in the XX century. By consequence, Buffon does not consider a species a set of individuals linked together by relations of resemblance but a genealogy of individuals – that is, historical successions held together by reproductive relations[11]. This view was embraced by Diderot in the *Encyclopédie* and adopted by Kant[12]. This emphasis on genealogy, rather than on the similarities among the members of a certain species, paved the way for a new conception of biological species no more regarded as abstract entities instantiated by singular organisms but as historical and complex entities formed by genealogical chains of organisms. The door was open for a different view of species not considered as abstract entities but as individuals.

2 Species as individuals

The concept of species that Darwin held has been vigorously debated for a long time[13]. Whatever it was, his evolutionary theory played a seminal role in the change of the ontological paradigm regarding biological species in the second half of the XX century. By introducing the concepts of mutation and natural selection, Darwin, in the first place, highlights new possibilities of evolution for species that previous naturalists had not discovered, limiting them to hybridization. In the second place, he makes it clear that evolution does not follow predictable lines that can be inferred from existing individuals. In the third place, Darwin claims that evolution can occur gradually over very long periods. The evolution of species does not necessarily occur only instantaneously, as in the case of hybridization, but also step by step and in a so slow way that cannot be appreciated by observing only a few generations.

According to the naturalists previous to Darwin, the offspring of a couple of organisms belonging to a species can diverge from their ancestors only by

10 Cf. Lovejoy (1968).

11 Cf. Sloan (1979).

12 Sloan (1979), pp. 127–128.

13 I will not attempt to outline even the general lines of this debate here; I refer to Stamos (2007).

contingent characters. Essential and defining characters must remain the same through generations. In Darwin's evolutionary view, this is not true and, in principle, every character of an organism can disappear in its offspring. This raises serious troubles for the classification of individuals and for their division into species. We can arrange organisms on a continuum along which they very gradually modify their traits; therefore, it is difficult to divide them sharply and to say unequivocally where the organisms belonging to a species finish and the organisms belonging to another begin on the continuum. Division in species seems to be arbitrary. Moreover, the distinction between defining and contingent characters disappears: every character can change, and none is necessarily constant through generations. When a character changes, it is difficult to decide whether it was a defining character of the species, so that the species evolved into another one, or whether it was contingent, so that the species remained the same. Every character is not stable in principle; therefore, the distinction between defining and contingent characters is difficult to trace.

How does one react to this state of affairs? A conventionalist perspective can be assumed by affirming that every distinction we trace in the biological world is arbitrary and conventional[14]. However, this is not the prevalent opinion among biologists and philosophers of biology, who are usually inclined to state the reality of species. Therefore, some alternative definitions of biological species have been formulated, and these are not grounded on the *characters* possessed by singular organisms – because they can vary in an almost limitless way – but on the *relations* among the members of the same species. In particular, two kinds of relations have been the focus: interbreeding relations and genealogical relations among organisms. Interbreeding relations are at the basis of the biological concept of species, which is probably, the most popular concept of biological species between the 1930s and the 1970s. According to this concept, a species is a group of individuals that can breed together and are reproductively isolated from other organisms[15]. The gene flow among the members of the same species and the lack of gene flow with the members of other species contribute to preserving the genetic pool of the species as relatively constant by recombining the genes of deviant individuals with those of conspecific individuals and by protecting the genetic pool from the introgression of genes belonging to individuals of other species. In this way, the most favorable combinations of genes are preserved,

14 For conventionalist views, see Vrana and Wheeler (1992) and Hendry (2000) and, for some aspects, see also Hey (2001).

15 For the modern biological concept of species, see Dobzhansky (1935), Dobzhansky (1937), Mayr (1942), Mayr (1949), and Mayr (1970).

and the production of too great a number of disharmonious, incompatible gene combinations is prevented.

In this framework, speciation is accounted for as the reproductive isolation of a population of organisms from the other members of the species due, for instance, to natural barriers. The consequence of such physical isolation turns out to be the separate evolution of the isolated population, which does not recombine its genetic pool with that of conspecific members. As a result of this independent evolution, the genetic pool of the isolated population becomes deviant compared to that of other members of the species, so that, even though the physical barrier is removed, the two populations can no longer interbreed. Since they became two reproductively isolated communities, they belong to two different species.

The drawbacks of the biological species concept – not the least of which is the fact that the concept can be applied only to sexual species, while most existing species are uniparental – had the effect that since the 1970s, biologists are mostly inclined to a different concept of species that is based not on interbreeding relations but on descent relations. The basic idea is that an individual is a member of the species *S* only if it descends from another individual of *S*. However, we cannot speak of a unique phylogenetic concept of species but rather of a family of concepts related to each other but having remarkable differences. In particular, a phylogenetic concept is incomplete unless it is not specified when a speciation episode occurs – that is, when an individual descending from a member of *S* is no longer a member of *S*. If this is not specified, the definition has the consequence that every descendant of an individual belonging to *S* is a member of *S*, and thus, that speciation is impossible. In this framework, the arising of a new species can be indicated either by the reproductive isolation of a community of individuals from the other descending members[16] or by the appearance of a new particular trait that is not present in the ancestors of an individual[17].

In some cases, phylogenetic concepts are based on more abstract criteria, which are sometimes difficult to interpret, such as those of G. Simpson, who defines a species as "a lineage (an ancestral-descendant sequence of populations) evolving separately from others and with its own unitary evolutionary roles and

16 Therefore, this view mixes elements of the phylogenetic concept of species with elements of the biological concept. Cf. Henning (2001).

17 These phylogenetic concepts, which are grounded on traits that are not present in the ancestors of a population, mix elements of the phylogenetic concept of species with elements of the classical morphological concept, as a set of individuals that instantiate some properties. Cf. Rosen (1978), Eldredge and Cracraft (1980), and Nelson and Plamick (1981).

tendencies"[18], or those of E. Wiley, who defines a species as "a single lineage of ancestor–descendant populations which maintains its identity from other lineages and which has its own evolutionary tendencies and historical fate"[19]. Of course, until there is clarification regarding what is intended by "evolutionary tendencies" and "historical fate", these definitions risk being generic and vague.

However that may be, everyone agrees that the phylogenetic criterion is not a sufficient condition for the definition of species. All living organisms likely descend from a unique organism or from a few organisms that lived some billions of years ago, but they obviously do not belong to the same species. Moreover, it can be questioned whether the phylogenetic criterion is a necessary condition for belonging to a species. Suppose, for example, that an organism that has both the genotype and the phenotype of lions is produced in a laboratory by means of a very advanced technique. According to the phylogenetic criterion, such an organism is not a lion because it does not descend from a member of the species of lions even though this organism is completely indistinguishable from other lions due to its features. This consequence is very difficult to accept.

However, it is not my aim to criticize the concepts of species grounded on the relations among the organisms rather than on their characters. It will suffice to say that these concepts have favored a new view of species that are not regarded as abstract kinds instantiated by singular organisms but as complex individuals composed of singular organisms. As Ghiselin (1974) argues, species are similar to firms. To decide whether a person is employed to a certain firm, we do not have to observe his or her features but the relations between him or her and the other employees of that firm. By consequence, a firm is not a species instantiated by its employees but a complex individual whose employees are parts linked by some relations. Indeed, the unity of the firm is due to the relations among its employees. The same can be said about species. In order to evaluate whether an individual belongs to a biological species, its features are merely an indication. What is decisive is the relations between the individual and the other members of the species. It is, therefore, natural to consider a species as a complex individual whose cohesion is determined by the relations among its members. Besides firms, biological species can be compared to other complex individuals whose unity is due to the relations among their members, such as musical bands or families.

However, the biological and phylogenetic concepts of species do not have the species-as-individuals view as a *necessary* consequence. In fact, the

18 Simpson (1961), p. 153.
19 Wiley (1981), p. 25.

species-as-kinds views that have relations among members as a criterion of membership are also possible[20]. To be sure, if the membership criterion is based on the relation among members, it is *more natural* to consider species as complex individuals whose cohesion is a result of the relations among the organisms that are parts of the species[21]. Moreover, given the evolution and the change of species, the difficulty in finding abstract traits that are common to every member of the species has the consequence that relational criteria are generally preferred[22].

If relational criteria suggest but do not imply the view of species as individuals, there is another argument that seems to have the species-as-individuals view as a necessary consequence: species have a beginning in time, an evolution, and an end. They seem to be temporal entities. Abstract entities are, instead, usually conceived of as timeless. Therefore, biological species cannot be abstract entities and, thus, cannot be kinds. Therefore, they are individuals[23]. The evolution of species seems to be a decisive reason for considering them complex individuals rather than kinds.

3 Evolving abstract entities

However convincing this argument may appear, it is not decisive. In fact, there are plenty of entities, almost unanimously considered abstract, that evolve or, at least, that we say evolve and change: languages, theories, cultures etc. These entities appear to have a beginning in time, an evolution, and an end, exactly as biological species. But this does not prevent us from considering them abstract. Therefore, the evolution of biological species cannot be a decisive argument for considering them individuals. Here, I will focus in particular on languages in

20 The basic idea is that the essential property for belonging to the kind *K* is to have a certain relation with an individual *i*. For such views, see Okasha (2002) and LaPorte (2004).

21 For the thesis that the relational criteria for membership of a species naturally lead to the view of species as individuals, cf. Crane (2004).

22 Some of the scholars who have expressed their preference for the thesis that species are individuals are Ghiselin (1966), Ghiselin (1974), Hull (1976), Hull (1978), Holsinger (1984), Ghiselin (1987), and Coleman and Wiley (2001).

23 For these arguments, see Hull (1978), "[species] are the entities which evolve as a result of selection at lower levels (…) Species as the results of selection are necessarily lineages, not sets of similar organisms" (p. 343). See also Stamos (2003): "on the modern view species are supposed to evolve. But if a species is an abstraction, or if it has an abstraction as one of its properties, then it cannot evolve" (p. 179) and Kunz (2012): "The idea of a biological class in the Aristotelian and Linnaean sense as existing in reality and the reality of organisms being subject to evolution, in a state of constant change with regard to their traits, are incompatible with each other" (p. 22).

light of both my philosophical education and the fact that biological species and languages show a relevant resemblance, which was noticed by Darwin himself:

> The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously the same [...] The frequent presence of rudiments, both in languages and in species, is still more remarkable [...] Languages, like organic beings, can be classed in groups under groups; and they can be classed either naturally according to descent, or artificially by other characters. Dominant languages and dialects spread widely and lead to the extinction of other tongues [...] The same language never has two birth places. Distinct languages may be crossed or blended together. We see variability in every tongue, and new words are continually cropping up; but as there is a limit to the powers of memory, single words, like whole languages, gradually become extinct [...] The survival or preservation of certain favoured words in the struggle for existence is natural selection. Darwin (1871), pp. 59–61.

We can summarize these observations in the following points:

- Languages, as species, have a beginning in time, evolve, and become extinct. In languages, new words are continuously introduced while others fall into disuse. Two or more words of similar meaning can compete, and one of them can replace the others. In the same way, members of a species compete with each other, and some traits can survive to the detriment of others.
- By evolving, a language can result in another language (anagenesis) or can divide itself into two or more languages that descend from the first one (cladogenesis). Therefore, the genealogical trees of languages can be reconstructed as genealogical trees of species are reconstructed. Comparing the two kinds of trees can be fruitful and can lead to a more precise reconstruction of them both.[24]
- Languages have internal varieties (dialects). Similarly, species often have varieties. It is sometimes difficult to establish whether two idioms are dialects of the same language or two different languages, because criteria can conflict with each other or be vague. In the same way, it is sometimes difficult to establish whether two populations are two varieties of the same species or two different species. In these cases, too, the criteria are often vague and in conflict.
- In some spatial regions, two languages can be contiguous and shade one into other. For instance, there can be dialects that are intermediate between two languages or varieties that have some of the characters of each of the two languages. Likewise, two species can crossbreed and generate varieties and species that are intermediate between the two.

24 For these kinds of attempts, see Cavalli-Sforza (2000).

- Linguists reconstruct the genealogy of languages by observing their features. Biologists do the same by observing the features of the members of different species.
- Synchronic and diachronic conceptions of languages are possible. Diachronic conceptions are committed to the study of the development of languages in time and to the reconstruction of the genealogical relations among languages. Instead, synchronic linguistics studies linguistic systems in certain temporal sections and the relations among words at those times. Likewise, synchronic and diachronic conceptions of species are possible. Diachronic conceptions relate to the reconstruction of the slice of the genealogical tree in which a species is located, while synchronic conceptions relate to the study of the traits and the behavior of the members of a species at a certain time[25]

It is clear, however, that languages are complex abstract entities. A language is constituted by two components: a) a system of signs and b) a set of syntactic rules. More in detail:

1. The first component of languages is a system of signs that have meaning (the lexicon of a language). The term "system" is often used because there is no language formed by a single sign and because the signs of language (the words) have a number of relationships of different kinds: phonetic, morphologic, grammatical, and semantic. Consider, for instance, the class of words having the suffix *-ful* in English (*useful*, *forgetful*, *hopeful*, etc.) or the class of common nouns. For the semantic relations, consider, for instance, the synonyms, hypernyms, and hyponyms or the words belonging to the same semantic field.
2. Syntactic rules allow for the merging of words to create more complex linguistic structures, such as phrases, sentences, and texts. Not every combination of words is permitted (consider, for example, the string: **the go with thus because*), and syntactic rules establish what combinations are allowed.

Linguistic signs are types that are instantiated in the tokens that the speakers use. When we make reference to the signs of a language, we clearly mean the types and not the tokens. Therefore, signs are abstract entities, and the first component of a language is constituted by a set of abstract entities. But the second component is abstract, too, being constituted by a set of rules. By consequence, languages are complex abstract entities that are formed by simpler abstract entities. Similar

25 For the diachronic and synchronic conceptions of species and for a comparison with the analogous linguistic conceptions, see Stamos (2002).

considerations could be extended to theories or cultures, but this issue will not be discussed here.

4 How can abstract entities evolve?

The fact that languages – that is, abstract entities that have several similarities to biological species – evolve is significant evidence that the argument against the species-as-kinds view based on their evolution does not succeed. However, a question is in order: how can timeless entities evolve and change? Timeless entities should not be able to change, because change requires time (in fact, based on some views, temporal entities are in time just because they change). How can languages change if they are timeless entities? In this section, it will be shown in which sense languages evolve and how this sense can be extended to biological species.

The types included in the system of signs of a language during a certain period are the types that are instantiated by the speakers of that language. In more common terms, the words that constitute a language at a certain period of time are the words that are employed by the speakers of that language in that period. The words that are no longer used, the words that are used by other linguistic communities, and the possible words that are not used by any linguistic community are clearly not part of a certain language *l* at a certain time *t*. The principle for including a sign in the sign system of *l* at *t* is to have a certain number of instances (tokens) at *t*. Of course, the precise number of instances required for inclusion in the sign system of *l* is vague. A word does not become a word of a language *l* simply because a single speaker of *l* uses it, nor is it a word of *l* even though a restricted number of speakers occasionally use it. Lexicographers must decide case by case whether to include a word in the dictionary of *l*, and their choices could be questionable and not be adopted by other lexicographers. However, lexicographers usually agree on the words of *l*. For instance, the word *heat* is certainly part of the lexical system of English, while the word *Zeitschrift* is not and is a part of another lexical system. Except in extreme cases, lexicographers agree about the system of signs of a language *l* at *t*.

In parallel, the syntactic rules that are part of *l* at *t* are the rules that are employed by the speakers of *l* at *t* – that is, the rules that are instantiated at that time. In this case too, the number of instances necessary for inclusion in the grammar of *l* is not completely clear. Again, the fact that a single speaker of *l* uses a certain construction is not sufficient for including that construction in the grammar of *l*. However, except in extreme cases, linguists usually agree about the

rules that are part of *l* and can codify them in grammar books. To sum up, we can say that a language *l* at a time t is constituted by the signs and the rules that are instantiated by *the speakers* of *l* at *t*, leaving vague the required number of instances.

Now, suppose that a certain word *w* of *l* falls into disuse and is no longer employed by the speakers of *l* at time *t'*. The sign system of l will no longer include *w* among its types. How does one react to this *situation*? There are two possibilities: either by affirming that, because the system changed, the speakers speak a language that differs from *l* or affirming that they still speak *l* even if *l* has changed a little. In the first case, we have two languages: the language *l* spoken at *t* and the language *l'* spoken at *t'*. These languages differ just because *w* is included in the sign system of *l* but not in the sign system of *l'*. In the second case, we have a single language - *l* - that changes because it includes *w* in its sign system at *t* but not at *t'*. We usually adopt the second way of considering this matter: we would not think that English is no longer spoken because a single word is no longer used. We would not say that a language that is different from English is spoken. Rather, we would say that English has changed a little bit—that is, has evolved. Similarly, although more stable than the lexicon of a language, syntactic rules can fall into disuse and be replaced by other rules. We would not say, however, that a language is no longer spoken if a single grammar rule has changed. Rather, we would say that the grammar of a single language has changed a little.

Notice, however, that both signs and grammar rules are abstract and timeless objects that do not change over time. What changes is the fact that a word has instances or that a rule is in use: the instances of words and rules are temporal, but the words and rules are not. The fact that a language changes does not imply that the types of that language change. The temporal aspect of a language does not descend from the types that compose it but from the fact that the types included in the sign and grammar systems of a language at a time *t* are those that have instances at *t*. So, what changes are the tokens, not the types.

Suppose that the changes in the lexicon of a language increases, for example, as a result of the introduction of many foreign words and the replacement of many existing words. Moreover, suppose that the grammar of the language also undergoes important modifications. Is the language still the same, or has a new language arisen? There is a great deal of vagueness here and many intermediate cases in which a decision is impossible. However, if the changes are substantial and broad, then we would no longer say that it is still the same language with modifications but rather that a new language has arisen.

Can these considerations be extended to biological species? I believe so. Suppose that biological species are complex systems of abstract traits. On this

view, a species is not defined by a single trait, but by a system of phenotypic and genotypic traits[26]. The traits that should be included in the definition of a species are the traits that are instantiated by the members of the species. If one or more populations of organisms instantiate a set of traits *S*, then the species will be constituted by *S*. As in the case of languages, there might be some problems regarding which traits are included in *S*. For example, the absence of a trait in only one member of a species is not sufficient for the exclusion of that trait from the set *S*. A substantial number of members of the species must fail to have that trait in order to exclude it from *S*. Conversely, if a trait is instantiated in a large number of members of the species, but not in every member, it could be necessary to introduce a name for a variety of organisms within the species—that is, the organisms that possess this trait. They constitute a race or a subspecies. Similar cases can be found with languages: if a considerable subset of speakers of *l* uses a certain specific lexicon that is partly different from that used by the other speakers of *l*, then linguists often identify a dialect within the language—that is, the dialect spoken by the speakers using that specific lexicon. If the differences between the organisms of *S* that exhibit these specific traits and the other members of *S* are numerous and significant, one can question whether they are simply a race within the species *S* and affirm that they belong to a different species. Likewise, if the differences in the lexicon that is employed by a subset of the speakers of *l* and that spoken by the other speakers of *l* are numerous and significant, one can question whether these speakers simply speak a dialect of *l* and affirm that they actually speak a language that is different from *l*. The question of whether two populations are two varieties of the same species or two different species is a normal subject of discussion among biologists.

Now, suppose that one of the traits *s* of *S* begins to fail to be instantiated in the members of the species, for example, as a result of mutated climatic conditions that tend to disfavor *s*. After a certain time, *s* is no more exhibited by the members of the species. Also, in this case, two reactions are possible: we can say either that a new species *S'* has arisen and that *S'* differs from *S* because of the lack of *s* or that *S* has evolved and changed due to the fact that *s* is no longer included in the set of traits *S*. In addition, in this case, we are inclined to react in the second way if *s* is just one of several traits included in *S*. Notice that in this case, too, the traits of *S* are abstract and timeless. What changes is the fact that some of these traits

26 For the sake of simplicity, I do not include relational traits, such as the possibility of interbreeding with other individuals of the same species, or the genealogical relations with other individuals of the same species. However, these relational characters can be included in the traits that define a species, even if this necessarily causes some complications.

are instantiated at different times. The organisms that instantiate the traits are historical entities, but the traits are not.

In parallel to languages, when the traits of S change in a massive way, it can be debated whether the evolved organisms belong to *S* or to a new species. Biologists can and often do disagree on this point. However, it is plain that when the changes are several and very significant, there is no doubt that *S* is extinct and that the evolved individuals *belong* to a species that is different from *S*. Both anagenesis and cladogenesis are allowed in this framework, as they occur regularly in linguistic evolution.

5 Conclusion

The strongest argument in favour of the species-as-individuals view is based on the fact that species have a temporal beginning, evolve, and become extinct, while abstract entities are timeless. However, the comparison with languages shows that this argument is much less strong than it might initially appear. In fact, it is possible to affirm that biological species are historical even though they are constituted by a set of abstract traits. As seen in Section 2, there are other reasons, although none decisive, for adopting the species-as-individuals view, grounded on the fact that for membership of a species, relational rather than intrinsic criteria seem to be central. Arguments of this kind depend on the success of the biological and phylogenetic definitions of species at the expense of the phenotypic and genotypic definitions. In fact, arguments can be advanced that aim to show that biological and phylogenetic definitions are insufficient and that phenotypic and genotypic traits cannot be neglected in a definition of species. However, this is a topic for another paper.

Maria Carla Galavotti

Who is Afraid of Subjective Probability?

Abstract: Although accredited by a great many people across a wide range of fields, the subjective interpretation of probability is still the target of skepticism and bitter criticism. In particular, it is accused of being unable to account for the objective meaning usually attached to probability in science, and of neglecting evidence that could support such a meaning. Its adoption in other contexts, for instance as applied in courts of justice, has also been opposed for similar reasons. This paper argues that criticism of subjective probability is largely motivated by misunderstanding. To support this claim, attention will be called to some aspects of Bruno de Finetti's viewpoint that have by and large been neglected by the literature, and to Frank Plumpton Ramsey's way of accounting for objective probability within the subjective theory, which has not received much attention either.

1 Subjectivism under attack

To start with, some criticism moved against the subjective theory of probability by authors working in different fields will be recalled.

In a "Technical report" on a sensitive subject, namely the chance of California being hit by an earthquake before 2030, statisticians David Freedman and Philip Stark criticize the subjective interpretation of probability – which they call 'Bayesian', following a currently widespread but disputable tendency because the two categories do not overlap since many Bayesians do not embrace the subjective theory. They write that "According to Bayesians, probability means degree of belief. This is measured on a scale running from 0 to 1. An impossible event has probability 0; the probability of an event that is sure to happen equals 1. Different observers need not have the same beliefs, and differences among observers do not imply that anyone is wrong. The Bayesian approach, despite its virtues, changes the topic. For Bayesians, probability is a summary of an opinion, not something

The topic of this paper is reminiscent of a number of conversations with Sergio Galvan on the nature of probability.

Maria Carla Galavotti: Department of Philosophy and Communication, University of Bologna

DOI: 10.1515/9783110529494-011

inherent in the system being studied. If the USGS says 'there is chance 0.7 of at least one earthquake with magnitude 6.7 or greater in the Bay Area between 2000 and 2030', the USGS is merely reporting its corporate state of mind, and may not be saying anything about tectonics and seismicity. More generally, it is not clear why one observer should care about the opinion of another. The Bayesian approach therefore seems to be inadequate for interpreting earthquake forecasts" (Freedman and Stark (2003), pp. 2–3).[1]

In a recent publication, philosopher of science Guido Bacciagaluppi, commenting on Bruno de Finetti's subjectivism writes: "Exchangeability is an assumption about the structure of one's subjective priors, and is entirely independent of any objective behaviour of the system under consideration. [...] according to de Finetti there is no sense in which our probability judgments are right or wrong. As de Finetti very graphically expresses, PROBABILITY DOES NOT EXIST" (Bacciagaluppi (2014), pp. 402–405).

Criminal law expert Mike Redmayne argues that the assessment of evidence in court needs objective probability, and discards subjectivism because "when the only constraint on rational belief is coherence among a belief set, it can seem that anything goes" (Redmayne (2003), p. 276).

In a similar vein, philosopher of science Larry Laudan, author of a number of writings on epistemological issues concerning criminal trials, objects to the use of Bayes' method in court on the grounds that prior probabilities are the expression of the "subjective hunches" of those who fix them, which would lead to the admission of arbitrary and discretional probability judgments.[2]

In a nutshell, the major charge moved against subjectivism is that it involves probability evaluations that are merely the expression of personal beliefs, and disregard empirical, 'objective' information.

2 Subjectivism vindicated[3]

As mentioned by Bacciagaluppi in the above quoted passage, Bruno de Finetti maintained that "*probability does not exist*", and wanted this sentence printed in

1 More criticism on the 'Bayesian' approach can be found in Freedman (1995).

2 See Laudan (2003) and Laudan (2006). More comments on the criticism moved by Laudan and other authors to the adoption of subjective probability in the context of criminal Law can be found in Galavotti (2012).

3 This section draws passages from Galavotti (2001) and Galavotti (2009).

capital letters in the *Preface* to the English edition of his *Theory of Probability*.[4] No doubt, his insistence upon this claim has been an obstacle to the diffusion of the subjective interpretation among scientists and other people operating in fields permeated with the conviction that probability should have an objective meaning. Nevertheless, de Finetti's claim should not be taken to mean that once coherence is satisfied probability can take on any value, nor that for a subjectivist there is no problem of objectivity of probability evaluations. Surely this is not what de Finetti meant. De Finetti rejected the "distortion" of "identifying objectivity and objectivism" (de Finetti (1962b), p. 344), but did not deny that evaluations of probability should obey a criterion of objectivity. In other words: de Finetti did not reject *objectivity*, but *objectivism*, namely the idea that probability depends *entirely* on some aspects of reality and is to be *uniquely* determined by evidence, an idea that he deemed "a dangerous mirage".

Albeit for de Finetti objective probability (uniquely determined, inherent to facts) does not exist, there is a *problem of objectivity*, namely the problem of devising "good probability appraisers".[5] In order to satisfy the criterion of objectivity, probability should be evaluated taking into account *all available evidence*, including frequencies and symmetries. That said, de Finetti thought it would be a mistake to put these elements, which are useful ingredients of the *evaluation* of probability and when available should not be ignored, at the basis of its *definition*. This kind of mistake, in his opinion, is made by the upholders of those interpretations that define probability only in terms of frequency or symmetry, namely frequentism, logicism and the classical approach. By contrast, de Finetti exhorts us not to conflate the definition of probability with its evaluation.

Having distinguished between definition and evaluation, de Finetti clarifies what he means by the evaluation of probability, emphasizing that the process through which probability judgments are obtained is more complex than is assumed by the other interpretations of probability, which define probability on the basis of a unique criterion. The choice of one particular function (among those satisfying coherence) is the result of a complex procedure, which involves context-dependent, and therefore subjective elements. To be sure, every evaluation of probability results from "the conjunction of both objective and subjective elements at our disposal" (de Finetti, 1973, p. 366). Far from neglecting objective information, de Finetti reaffirms that both "(1) the objective component,

4 See de Finetti (1970), English edition 1975.

5 The expression is borrowed from the Bayesian statistician I.J. Good. See Good (1965) and Good et al. (1962).

consisting of the evidence of known data and facts; and (2) the subjective component, consisting of the opinion concerning unknown facts based on known evidence" (de Finetti, 1974a, p. 7) are essential. To sum up, according to de Finetti "subjectivism is one's degree of belief in an outcome, based on an evaluation making the best use of all the information available to him and his own skill. [...] Subjectivists [...] believe that every evaluation of probability is based on available information, including objective data" (de Finetti, 1974b, p. 16). These passages should be sufficient to absolve de Finetti from the charge of overlooking objective information, and to consider equally acceptable whatever evaluation is made.

Against the habit of relying on ready-made recipes applicable to all situations, de Finetti calls attention to the context-dependent character of both the objective and subjective components of probability evaluations. He invites us to consider that factual evidence must be collected carefully and skillfully, and that its exploitation depends on what elements are deemed relevant to the problem under consideration. Furthermore, the collection and exploitation of evidence depends on a number of contextual considerations, often of economic nature, but of other kinds as well. In an array of situations ranging from medical diagnosis and prognosis to weather and economic forecasts, it is hard to deny that personal experience and the degree of evaluator expertise play a crucial role on how to weigh experiential data.

Although the intrinsic context sensitivity of probability evaluation makes the idea of *absolute objectivity* nonsensical and, as pointed out by Bacciagaluppi, probability judgments cannot be deemed right or wrong, for subjectivists it is perfectly sensible to regard some evaluations as better than others. For that purpose they adopt a number of methods devised for improving probability evaluations, which are the object of a vast literature to which de Finetti has made a substantial contribution, partly in collaboration with Leonard Jimmie Savage. Their approach revolves around penalty methods like Brier's rule, named after the meteorologist who applied it to weather forecasts.[6] Methods of this kind are perfectly legitimate and justifiable within a subjectivist framework: as de Finetti emphasized: "though maintaining the subjectivist idea that no fact can prove or disprove belief, I find no difficulty in admitting that any form of comparison between probability evaluations (of myself, of other people) and actual events may be an element influencing my further judgment, of the same status as any other kind of information" (de Finetti (1962a), p. 360).

The preceding considerations show that the claim that the subjective interpretation of probability involves a commitment to an "anything goes" approach

6 For more on this see Dawid and Galavotti (2009).

is ill-founded. That said, it is undeniable that de Finetti was more concerned with the application of probability to everyday life, and more specifically to those situations where only scant information is available, than to science. In the pragmatist spirit that imbues his perspective, de Finetti regarded science as a continuation of everyday life, and held that subjective probability is perfectly adequate for both. His attitude is rooted in the result known as "de Finetti's representation theorem", which shows that with increasing evidence the adoption of Bayes' rule in conjunction with exchangeability guarantees convergence between subjective degrees of belief and observed frequencies.[7]

Although de Finetti did not explicitly address the issue of defining a notion of "probability in science", Frank Ramsey did, coming to an insightful way of accommodating the notions of "chance" and "probability in physics" within the subjective outlook.

3 Subjectivism and probability in science[8]

Ramsey's insightful way of accounting for the notions of *chance* and *probability in physics* deserves more attention than it has received in the literature. First of all, Ramsey defines chance as degree of belief of a special kind. Its peculiarity is that of being always referred to a *system of beliefs* rather than to the beliefs entertained by single agents acting in certain situations. The systems to which chance is referred peculiarly include laws and other statements describing the behaviour of the phenomena under consideration, such as correlation statements. Taken in conjunction with the empirical knowledge possessed by the users of the system, such laws entail degrees of belief representing chances, to which the beliefs of different agents should approximate. This notion typically applies to chance phenomena like games of chance, whose behaviour is described by systems that cannot be modified by the addition of deterministic laws ruling the occurrence and non-occurrence of a given phenomenon.

On the basis of this concept of chance, Ramsey defines *probability in physics* as chance referred to a more complex system, namely a system making reference to scientific theories. In other words, probabilities occurring in physics are derived from physical theories, and can be taken as *ultimate chances* to mean that within

7 This result, which was found by de Finetti as early as 1928, is spelled out in some detail in de Finetti (1937).

8 This section draws passages from Galavotti (1995) and Galavotti (1999).

the theoretical framework in which they occur there is no way of replacing them with deterministic laws. Chances derive their objective character from the theories which are "taken as true" because they meet with the consensus of the scientific community. In this connection it is worth noticing that within Ramsey's perspective the truth of theories is accounted for in pragmatic terms. In fact Ramsey holds the view, whose paternity is usually ascribed to Charles Sanders Peirce, that theories which attain "universal assent" in the long run are eventually accepted by the scientific community and taken as true. Along similar lines, Ramsey characterizes a "true scientific system" with reference to a system to which the opinions of everyone, grounded on experimental evidence, are bound to converge.[9]

Plainly, Ramsey's suggestion makes it is perfectly possible to accommodate within the subjective outlook a notion of probability endowed with the objective character that is required by science. The idea that probability as degree of belief can be guided and even determined by scientific theories is fully compatible with de Finetti's viewpoint. As a matter of fact, the posthumous volume *Filosofia della probabilità* contains some remarks admitting that probability distributions belonging to scientific theories can be taken as "more solid grounds for subjective opinions" (de Finetti (1995), English edition 2008, p. 63). As an example, de Finetti mentions statistical mechanics. Such remarks provide evidence that late in his life de Finetti entertained the idea that probabilities occurring in physics are endowed with a peculiar robustness which derives from the theories describing the behavior of the phenomena under consideration, and explain how their probability should be fixed.[10]

4 Conclusion

The preceding remarks aimed to show that the subjective theory of probability is quite different from the way in which it is often depicted. Far from representing an anything goes approach, subjectivism has the resources to distinguish between good and bad ways of evaluating probability, and also to account for the notion of objective probability. While such resources can already be found in the version of subjectivism elaborated by its principal founders, namely Ramsey and de Finetti,

9 See Ramsey "Truth and Probability" and "Chance", both written in 1928 and published in Ramsey (1931), "General Propositions and Causality" also in Ramsey (1931) reprinted in Ramsey (1990). See also Ramsey (1991), notes 67 and 69.

10 This is argued in some detail in Galavotti (2001) and Galavotti (2005).

more recent work by a long list of authors such as Richard Jeffrey, Patrick Suppes, Brian Skyrms, Wolfgang Spohn, to mention but a few, has more to offer to those who are not afraid of the term "subjectivism". Granted that the term is not a happy one – Savage proposed substituting it with "personalism", but his suggestion has not won much acclaim – one should nonetheless take the trouble to understand what it has to offer.

Georg Gasser

Agent-causation and Its Place in Nature

1 Doubting Humeanism

Many contemporary philosophers are consciously or unconsciously in the grip of a Humean conception of reality. A Humean conception of reality claims that all that exists, ultimately speaking, are basic entities of a specific type (events, for instance) and relations between them. Relations are described as causal if their relata succeed temporally, are contiguous, are qualitatively similar and follow a repetitive pattern. For causation to happen, no further ontological ingredient is required. One can dub such a conception of reality "structuralist" and "actualist": "structuralist" because the world is ultimately a complex structure of spatiotemporal relations depending on the specific distribution of basic entities, and "actualist" because the basic entities are locally instantiated qualities. Thus, to use Hume's famous phrase, all ultimate entities seem entirely loose and separate; they seem conjoined but never connected.(Hume (1999), 7.2.26.) A prominent contemporary defender of this view is David Lewis. According to him,

> all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another. (Lewis (1999), ix)

Such a view is, ontologically speaking, extremely sparse, which makes it attractive. It complies with the researcher's old dream to explain nature's complexity in strikingly simple terms. Take, for instance, the observation that salt dissolves in water. Explaining this observation requires no recourse to a substance, salt, or its special dispositional properties, such as being water-soluble; to account for the electrostatic process of hydration, all that we need to do is point to the physico-chemical properties of salt and water.

Among the major motives of philosophers defending a Humean account is the worry that reference to dispositions and powers unnecessarily conflates our ontology and, even worse, opens the door to entities which science cannot grasp. The vase's disposition to break appears invisible; all that we can see is that the vase breaks when struck with a hammer. Salt's disposition to dissolve in water is unobservable; all that we can observe is that the grains of salt have dissolved after being placed in a sufficient quantity of unsaturated water. Given that the physico-chemical structures of the entities involved in these causal processes

Georg Gasser: Innsbruck University

DOI: 10.1515/9783110529494-012

are open for empirical investigation, why suppose that there are dispositional properties? Lewis emphasises this view when he writes that

> [t]he point of defending Humean Supervenience is […] to resist philosophical arguments that there are more things in heaven and earth than physics has dreamt of. (Lewis (1999), p. 474)

Without going into further detail, it should be obvious that a conception of the world as ultimately consisting of loose and separate entities in certain structural relations to each other lacks the means for explaining why the world displays the kind of regularity that we experience. For the central point of this view is to deny that anything more robust *can* be said about the world's holding together and unfolding (rather) regularly over time. Thus Helen Beebee admits that

> if there really is nothing in virtue of which the universe is regular, then the fundamental nature of the universe is analogous to the story being played out on the computer screen: it's just a continuous fluke that things go on in the orderly way that they do. (Beebee (2006), p. 527)

From the perspective of common sense, this consequence of a Humean conception of reality seems hard to swallow. Galen Strawson, for instance, argues that part of a realistic outlook of reality is the view that material objects inhabiting this world can affect and modify each other in particular ways, and these events of affecting and modifying are constitutive of what we take causation to be.[1] The common-sense view of causation involves an element of production: If a grain of salt dissolves in water, then water molecules affect its crystalline grid in such a way as to break up the stable structure of the grid. The interactions between water molecules and molecules of the crystalline grid cause the grain to dissolve. Similarly, if a porcelain vase is struck by a hammer, then the force exerted by the hammer affects the fragile structure of the vase, thereby producing its breakage. If this production view of causation is taken seriously, then it apparently contradicts a Humean notion of causation.[2]

Strawson's critique goes a step farther. A mere production view of causation could also be true in a disordered world where entities act on each other arbitrarily. This, however, does not seem to be our world. Things appear to persist through time and to interact with each other in a regular fashion. The obvious

1 Strawson (1987). A structurally similar argument can be found in Esfeld (2007).

2 Of course, things get even worse once it comes to human action. Within a Humean metaphysics one cannot admit that our experience as agents in the world as being veridical. A similar argument as Strawson's can also be found in Esfeld (2007).

explanation for this observation, according to Strawson, is that things possess and retain certain properties throughout their existence. In virtue of their respective properties, some things can enter into certain forms of interaction with other things. Thus, the assumption that things dispose of a determinate nature is a more probable explanation of the persisting order of our world than the Humean indication that the world's regularity is all just a matter of luck.

Other arguments have recently been advanced which bolster Strawon's argument. Here is one: some philosophers of science point out that the ultimate structure of reality does not preclude dispositional properties. Take electrons, for instance: these don't seem to have an inner structure, but they have spin, negative charge, a magnetic moment, etc. These properties are best described as dispositional; and if electrons indeed are structureless, then these dispositions are not reducible to anything purely categorical. To put it another way, these dispositions are fundamental and ungrounded.[3] Although it is reasonable to assume that the dispositions of a complex object ? such as the fragility of a vase or the solubility of a grain of salt ? are based on the object's microphysical structure, it does not follow that all dispositions are reducible to non-dispositions. It follows only that a rather large number of dispositions of macro-sized objects is reducible to a small number of dispositions at the microphysical level. Molnar summarizes the discussion as follows:

> our best credentialled sources of empirical knowledge suggest, to a very high degree of probability, that there are no properties that could serve as putative bases for the powers of the fundamental constituents of the physical world. (Molnar (2006), p. 137)

Interestingly, David Lewis's description of nature's fundamental properties as

> perfect natural intrinsic qualities, or of point sized-occupants of points. (Lewis (1999), p. 226)

seems to fit the characterization of an entity's basic dispositions. If electrons lack an inner structure, as current physical theories suggest, then an electron's spin or charge is a (i) fundamental, (ii) natural, and (iii) intrinsic quality which (iv) is instantiated at a point. Nevertheless Lewis did not regard dispositions as belonging to the fundamental level of reality. One major reason for Lewis's rejection of dispositions appears to be his worry that metaphysical speculation might go far beyond what empirical evidence takes for granted; another is his commitment to Hume's denial of any necessary connections in nature. Both reasons may be questioned, however: regarding the first reason, contemporary

3 See Mumford (2007) who speaks of the ungrounded argument.

scientific theory speaks openly about the dispositional nature of ultimate physical particles. Regarding the second, Hume's metaphysical and epistemological reasons for denying any robust notion of causation are highly contestable. There is thus ample room for debate whether we should follow a Humean account of reality.

This brings me to the second argument for Strawson's view: contrary to Hume, we do directly experience causation. I consciously say "causation" instead of "necessary connection" because I think that Hume was right to claim that we are unable to see necessary connections. He was wrong, however, to identify causation with a necessary connection. Briefly put, here is why[4]: In order for a causal process to be truly necessary, cause *A* would have to produce the effect *B* under any circumstances. Causal processes take time, however, and it is therefore always conceivable that a causal process will at some point be interrupted or suspended by an interfering factor so that the effect does not come about. Examples of such cases are legion; nature's causal order is after all untidy.

If this argument is sound, then intermingling causation and necessity results in a distorted view of causation. Anscombe calls attention to this point in her observation that

> [e]ffects derive from, arise out, come of, their causes. […] Now analysis in terms of necessity or universality does not tell us of this derivedness of the effect; rather it forgets about that. For the necessity will be that of laws of nature; through it we shall be able to derive knowledge of the effect from knowledge of the cause, or vice versa, but that does not show us the cause as source of the effect. (Anscombe (1993), pp. 91–92)

In other words, necessity's natural home is the realm of logic; it is a relation holding between propositions. Causation, by contrast, holds between states of affairs in nature. Its modal force is less strong than necessity, for a cause can produce a corresponding effect but its doing so (because of possible interferences) is not necessary. Once the notion of necessity is separated from that of causation, we are in a better position to understand what it might possibly mean to experience causation.

Taking up the view that an entity can in virtue of its causal properties enter a certain spectrum of interactions with other entities, we can conceive of causation as a manifestation of the respective entities' causal properties. An entity disposes of a specific dispositional causal profile which is manifested if the right circumstances obtain. We are aware of this dispositional nature in many cases:

4 In Mumford and Anjum (2011), chap. 3., this argument is developed and defended in detail.

we know that we should not light a cigarette at a gas station, for gas has the disposition of being flammable. We know that we should not touch an electric cable, for electricity has the disposition of causing dangerous injuries. We know that a porcelain vase is fragile, so to protect it when shipping it we pack it in styrofoam.

But it is not just the way we think about these and similar situations which indicates that we are familiar with the dispositional causal profile of many material objects. A strong case can be made for the claim that we even have a more direct and immediate access to the world's causal powers. Consider the following cases: You are lost in thought and you bump into a door. You are lifting weights in the gym and you feel how the dumbbells exert a pull on your arms. You experience the effort it takes to cycle up on a mountain. I would like to argue that in these examples we experience the causal powers of various objects firsthand — the hardness of the door, the heaviness of the weights, and the gravity on the slope. Material objects resist our planned activities in various ways, and we feel the effort it takes to overcome this resistance. We even have a specific sensory system for perceiving these causal influences and for responding appropriately to them: proprioception, that is, one's capacity to track one's bodily location, posture and limb position. A propioceptory system enables us to register the causal influences imposed on us by a material world and to respond to it accordingly. Lifting weights requires a sense of the right amount of effort required: If you exercise too little force, you will not succeed in lifting the weights; if you use too much, you will throw the weights over your shoulder.

To sum up: The Humean claim that, metaphysically speaking, causation is nothing more than a constant conjunction of one entity next to another is highly disputable. I presented three arguments that create room for discussion. First, the Humean view is disputable from the perspective of a realist outlook on reality: it is hard to believe that the rather ordinary course of the world is just a matter of luck and that no metaphysical explanation accounting for this order is available. Second, the Humean view is disputable from the perspective of contemporary science: current theories of particle physics seem to suggest that the nature of the ultimate physical particles is dispositional rather than categorical. Third, it is disputable from the perspective of our experience: an argument can be made that as embodied beings we do indeed experience causation, because the materiality of the world often resists our actions and successful agency requires overcoming this resistance of the material world. From these arguments a further argument, appealing to coherence, may be constructed. Accepting a productive view of causation enables us to connect our commonsense view of reality with scientific findings about the world's physicality and our experience as embodied agents in a material world. Thus, a productive account of causation contributes to the

groundwork for a coherent view of reality ranging from physical particles to our "Lebenswelt".

2 Events, powers and substances

The previous section contrasted two general metaphysical pictures of reality. It is easy to see that a Humean account is inimical to the very concept of agent causation, for the Humean metaphysics lacks a robust causal understanding. It is thus unsurprising that agent causation is viewed with skepticism by those in the grip of Humean assumptions. In this paragraph I will undertake a closer examination of event causation. I will argue that, once an appropriate metaphysical framework is admitted, a case for reducing event causation to agent causation can be made.

If we consider everyday speech, then we notice that event- and agent-causal formulations are common. We say such things as "Gill caused the vase to break" and "Gill's hitting the vase caused it to break", or "The bomb caused great damage" and "The bomb's going off caused great damage". The grammatical subject of the verb "to cause" can be an animate human being such as Gill, or an inanimate artefact such as a bomb, or it can be a particular event. Since objects and events belong to different ontological categories, one may infer that ordinary parlance suggests the existence of two different forms of causation – causation by an agent[5] and causation by an event.

However, standard contemporary causal accounts suggest that agent-causation can be analysed semantically in terms of event-causation, where the former is ontologically reducible to the latter. The idea is as follows: A statement such as "Agent *A* caused event *e*" can be analysed in terms of

I "There is an event *x* which involves agent *A*, and event *x* caused event *e*." (Lowe (2008), p. 123)

The causal structure of many causative verbs appears to support this analysis. Consider verbs such as "kill", "put down", "stop", "rip", etc. We may say that these verbs convey the semantic meaning of a means-end-structure. "To kill" means to cause the death of a living being by the use of some means to be further specified; "to put down" means to change an object's position by ceasing to hold it; "to stop"

5 Following the Latin term "agens" I use it in a rather liberal way tantamount to the concept of substance. Thus, If I am talking about "agent causation" it comprises causation by all kinds of substances. There is no need to restrict this concept to animated or even only rational beings.

means to cause the cessation of an object's motion by exercising some force upon it, etc. In these cases a sentence of the structure "A killed B" can be translated event-causally in "A's killing caused B's death". However, this proposal is open to a number of objections.

First, it is widely assumed that certain actions, so-called basic actions, defy such an analysis. Basic actions are not brought about by the performing of some other action but are instead performed directly by the agent. Take, for instance, blinking. If I decide to blink immediately, then I blink immediately. The question by what means I was able to blink does not seem to have any reasonable answer. I may blink in order to do something else, such as giving you a previously agreed sign in a poker game. Blinking can thus be a means toward a further end, but in order to blink I don't have to perform any other action first. Pointing out that my eyelid muscles have to be moved properly for my blinking to take place is of no help here. The proper movements of the relevant body parts is a necessary bodily requirement for blinking, but because these movements are outside our conscious control, they can hardly be described as further actions. On the contrary, it seems more appropriate to say that, by blinking, we cause these movements as physiological realizers of the blinking.

Second, it is a mistake to assume that a noun-phrase referring to an object as the grammatical subject of the verb "to cause" is an elliptic form of speaking because the standard logical form contains a noun-phrase referring to a particular event.[6] Consider the following sentences:

- "The bomb destroyed the bridge" *vs*. "The exploding bomb caused the collapsing of the bridge"
- "John lighted a match" *vs*. "John's lightening the match caused the igniting of the match"
- "Gill opens a window" *vs*. "Gill's moving her hand caused the opening of the window".

Notice that the sentences on the left are not syntactically incomplete. Translating these sentences into an event-causal form as shown on the right does not imply adding anything that was previously missing. Rather, such a translation requires us to provide additional information about the events involved, for the verb "to cause" does not as such convey the depth of meaning of the verbs in question used in combination with the grammatical subject and object in the left-hand sentences. Generally, we know that a bomb destroys a bridge by causing it to collapse due to the force of the blast, and we are aware that lighting a match

6 For a detailed analysis of this and related issues see Keil (2000), pp. 373–383.

amounts to striking it on a rough surface until it ignites. Without this (implicit) knowledge, a desired translation in event-causal terms would fail. The last example complicates matters further. Though we know that a human person has to do something to open a window, it is hard to identify two events here. The movement of the person's hand simply consists in the opening of the window. It is unreasonable to claim that the movement of the hand takes place first and then this causes the opening of the window. A reformulation along the lines of "the moving of the hand happened in such a way that the window opens in the next instant" is at best a clumsy reformulation of the easily comprehensible sentence that a person opens a window with her hand (instead of with the electric window opener), but is never a more lucid explication of that sentence. It is more than disputable that, if someone opens a window, a second event takes place at all. Thus, sentences of the logical form on the left cannot automatically be transformed into sentences of the logical form on the right. The assumption that a sentence conveying causal information can be divided into two events, a prior cause and a subsequent effect, is more than doubtful.

Third, consider the following proposals for translating a sentence such as "By doing *X*, *A* causes *Y*" into "*A*'s *X*-ing caused *Y*'s *F*-ing" or "There is an event *x* which is an action of *A*, and event *x* caused event *y*". Although the proposed translations "*A*'s *X*-ing caused *Y*'s *F*-ing" and "There is an event *x* which is an action of *A*, and event *x* caused event *y*" appear structurally similar, there is a crucial difference between them. If I refer to *A*'s *X*-ing, I remain neutral about *A*'s role in causing an event. It could be that *A*'s *X*-ing happens automatically and so cannot be brought under any intentional action description. If I say instead that event *x* is an action of *A*'s, then *A*'s active role is explicitly confirmed because, for an action to take place, an agent is required who or which brings it about. It appears to be a clear category mistake to say that events can perform actions too. If I am right, then these considerations seem to vindicate the claim that agent-causation ought to be differentiated from event-causation. Consequently, the thing for a proponent of event causation to do is find a way to define the concept of action without any reference to an agent actively bringing it about. "*A*'s action" has to be analysed in event-causal terms with no allusion to the agent at all.

Uwe Meixner discusses two ways how this might be achieved (Meixner (2001), pp. 323–335). Either you aim to spell out a non-causal concept of agency, or you aim to describe the events causing the action as not being causally related to the agent herself. Consider the latter strategy first: if actions are events caused by

other events which cannot be causally attributed to the agent herself, then one must claim that

> the causal role assigned to the agent by common sense reduces to, or supervenes on, causal relations among events and states of affairs. (Velleman (2000), p. 130)

In other words, the level of action-involving events is transferred from an agential or personal to a sub-agential or sub-personal level. Causal mechanisms within the agent assume the role traditionally assigned to the agent herself. It is doubtful whether such a strategy is of any real help. A mere transference from the agential to the sub-agential level masks the original *analysandum* without providing any positive account of it. If all we say is that actions are not directly attributable to the agent herself but to action-initiating mechanisms within her, then, still, a positive account must be given of the sense in which this mechanism initiates an action rather than a mere reaction or reflex.

To put it another way: as long as actions are distinct kinds of events, something has to account for this difference. The obvious suggestion within a causal framework is to look at the causal history of actions for identifying the special causal ingredient which turns a causally "ordinary" event into an "action-event". This suggestion entails a distinction of two different forms of causation: events which cause action-events and events which cause ordinary events. Although we might be in a position to spell out such a distinction in event-causal terms only, a causal dichotomy still lurks which is structurally parallel to the distinction between agent causation and event causation. In either case a special concept of causation is invoked to explain the production of an action.

This brings me to the second strategy. The obvious way to circumvent the challenge of explicating the special causal history of actions is to abandon any form of agent-causal jargon altogether. One might say, for instance, that the causes of an action are within the agent herself but that no special action-initiating mechanisms are involved. To attribute an action to an agent, it suffices to show that the causes are within the agent. Once this approach is pursued, however, it is hard to see how the concept of action can be meaningfully attributed to the *analysandum* at all. As Irving Thalberg remarked many years ago, on such a view the agent mutates into a mere area where

> 'his' calculations, his perceptual judgments, his noble and base inclinations, perhaps his repressed fantasies, his conscious terrors, rages, lusts and devotions, either contend or blend with each other. Even if these proceedings do generate agitation of his limbs, why should we say that this is "his act". (Thalberg (1980), p. 220)

If the criterion for attributing an action to an agent is only its spatial relation, then we seem to be deprived of any means of identifying the specific difference between an intentionally executed action and a mere bodily motion happening to us. Both events may be "of the agent" in terms of taking place within the agent. But now we are left empty-handed. No proper concept of action is available anymore. The very precondition for differentiating a person's action from any other form of her bodily motion — that something is up to the agent — is discarded.

These reflections show that the project of reducing agent causation to event causation encounters severe problems on the semantic and ontological levels. If agents perform any actions at all, it is likely that they perform basic actions, and these are not explicable in terms of one event causing another. Rather, basic actions seem to be performed by an agent directly, which amounts to an instance of irreducible agent causation. The aim of reducing agent causation to something non-agential results ultimately in the annihilation of the very notions of agent and action. Agent causation is not reduced to but swallowed by event causation.

In light of these prospects, one might ask whether it might not be worthwhile to try to change the direction of analysis. Instead of explaining agent causation in terms of event causation, one could aim at explaining event causation in terms of agent causation:

II Event *x* caused event *e* if and only if there was some agent *A*, some manner of acting *F*, some agent *B*, and some manner of acting *G*, such that *x* consists in *A*'s *F*-ing and *A* by *F*-ing, caused *e*, which consisted in *B*'s *G*-ing.[7]

This analysis accords well with the outline of the dispositionalist metaphysics drawn in the first paragraph. Consider once more the statement that the exploding bomb causes the bridge to collapse. This statement is true because the following conditions are met: (i) There is a substance, the bomb, and part of its specific dispositional causal profile is to be explosive. (ii) Due to particular circumstances, this disposition was manifested; and its manifestation consists in the bomb's exploding. (iii) There is another substance, the bridge, disposing of another specific causal profile. This profile includes the disposition to collapse if a strong enough force acts upon it. (iv) The exploding of the bomb is a strong enough force, and therefore the bridge's disposition to collapse is manifested; this manifestation consists in the event of the bridge collapsing.

This account explains how the two events, the explosion of the bomb and the collapse of the bridge, are causally connected. The connection results from

7 This is an expanded proposal of the one suggested in Lowe (2008), p. 136.

substances which, in virtue of their properties, dispose of a causal profile which is manifested under certain conditions. Causation involves the exercise of different substances which act with and against each other in a variety of ways.[8] It is not hard to see that this account gives an important role to events in causation but at the same time implies that events themselves are causally impotent. Events do not cause other events to happen; rather, events are the result of particulars which interact with each other by manifesting their respective causal dispositions. If a particular's dispositional causal profile were not be manifested, then no event would be instantiated; the reason is that the underlying particular would be inactive. Surprisingly, therefore, this view accords with the Humean claim that events are causally impotent – although its reasons differ significantly from the Humean's. The Humean claim says that causal potency is not a basic feature of reality, for all that there is, are loose and separate events which succeed each other. The account defended here, by contrast, claims that events are causally inert because they are parasitic upon the causal workings of particulars.

Apart from the widespread acceptance of a broadly Humean framework, there is a further explanation for the predominance of event causal accounts which is suggested by Lowe.[9] He thinks that we tend to resort to this account when we are at least partially ignorant about the real powers at work in a given case of causation. We can claim that event *x* caused event *e* even though we are not in a position to explain in detail why this is so because we are unable to identify the specific substances acting upon each other. Events are epistemically more easily accessible than the underlying powerful substances, and this epistemic accessibility mistakenly involves the inclination to ascribe ontological primacy to events as well as substances. Mumford makes a similar observation when he writes that

> we need to distinguish a factive level of what happens in epistemically easy events from a transfactual level of powers that combine to produce those events. (Mumford (2009), p. 108)

The account proposed here holds that agent causation enjoys ontological and conceptual primacy, and it explains how events factor into this picture. Due to lack of space I leave more detailed explications for another occasion. My point is simply that a Humean causal account is by no means the only game in town. On the contrary, there are strong reasons favoring the alternative account of agent causation. I conclude this section by examining three influential objections put forward against it.

---

8 A detailed explication of this view – although in pandispositionalist terms – is found in Mumford (2009).

9 Lowe (2008), pp. 138–139.

2.1 The timing objection

The timing objection against agent causation says:

> How could an event possibly be determined to happen at a certain date if its total cause contained no factor to which the notion of date has any application? And how can the notion of date have any application to anything that is not an event? (Broad (1952), p. 215)

The kernel of this objection is that reference to a cause should explain why its effect occurred at a given time and not earlier or later. Pointing to a substance *holus bolus* does not provide such an explanation because a substance exists before and presumably also after the explanandum. What appears to be correct about Broad's objection is the claim that an entity's causing something results in that something's happening. A change calls for an answer to the question of what exactly effectuated this change now. However, as we have seen, this observation does not imply that agent causation ought to be substituted by event causation. It is one thing to claim that a particular can only cause by manifesting its dispositions and quite another thing to claim that events themselves are causes. If a substance disposes of a particular causal profile, then an event consists of a substance's causing something due to this profile. Thus, the event depends upon the substance's being causally active in one way or another. Consequently, events are not the right ontological category by which to account for direct causation, because it does not seem proper to say that events dispose of causal powers. This is not to deny that events take place when causation takes place. Rather, it is to deny that events themselves are the causal source bringing about the effect to be explained.

Note that this explication does not claim that a substance as such is the cause of a determinate effect. This claim would indeed be mysterious, because it is hard to see how a substance would manage to bring about an effect without acting in one way or another. Thus, a proponent of agent causation should not say that causation consists in a simple alteration of the event-causal model by replacing the event as cause with the agent as cause. The idea is not that one causal *relatum* is substituted with another.[10] Rather, the claim is that causation is less a relation between two separated entities than a manifestation of a substance's causal profile acting upon another substance (or substances). Causation is not a relation tying together two relata together nomologically, probabilistically or counterfactually; rather, it is the transition of a substance's causal state from

10 Clarke (2003), p. 186, makes such a suggestion when he writes that "when an agent causes an event, the relation in which the agent stands to that event is the very same one in which one event stands to another when the first causes the second."

potency to actuality, a transition which entails a range of effects upon other substances. These effects in turn manifest determinate effects depending on their specific causal profile.[11]

2.2 The objection of non-analysability

The objection of non-analysability claims that the agent's directly causing something remains mysterious because there is no way to analyse the causal relation between the agent and his or her causing. The fact that event causation allows for such an analysis whereas agent causation does not appears to be a crucial objection to the latter. However, we must bear in mind that the concept of the non-analysability is built into the very concept of agent causation. The notion of "directly causing" excludes any internal causal structure and therefore no further causal analysis is available.

Take the example of a radium-atom decaying spontaneously. If we understand that the atom's nature is to decay spontaneously and unpredictably, it is no longer relevant to ask why the nucleus decays now and not at some other moment. Any attempt to divide the event of decaying into two separate further events, one being the cause and the other the consecutive effect, is doomed to fail. All that can be said is that a spontaneous decay is the cause of several effects, such as the emitting of alpha particles and gamma rays. Similarly, one can claim that a spontaneous action is directly caused by an agent itself. Any further causal analysis will lead us down the wrong path, because we will begin to look for items inside the agent to give an action-explanation. This is not to ignore the major differences between an atom decaying suddenly, the neighbour's cat moving spontaneously and my directly bringing about the intention to finish this paper. However, what all these cases have in common is the same basic ontological structure of causation: in each case an entity is endowed with specific causal powers which enable it to produce a range of effects directly, so that no external causal trigger is required in the first place.

2.3 The impossibility objection

One may accept the answers to the two previous objections but advance this one instead: Although agent causation as such seems theoretically intelligible, there is

11 Mumford (2009) defends a pandispositionalist version of this account. He sees causation as a shifting around of different powers as his title of the paper indicates.

a strong reason why it is impossible. You, the proponent of agent causation, claim that substances have certain causal powers in virtue of having certain properties. If a substance's properties were different, then its causal profile would be different too. Consequently, the real cause of the substance's behavior seems to be that substance's having its intrinsic properties, not the substance itself.[12]

This objection works on the basis of a metaphysical distinction between powerful properties possessed by a substance, on the one hand, and the (powerless) substance itself, on the other. Someone may limit the ontological function of a substance to being merely the bearer of powerful properties. Then the substance (or agent) itself vanishes from the causal picture of reality – for only the substances's intrinsic properties relate causally to each other. It is disputable, however, whether the distinction between a substance and its properties amounts to a *distinctio realis*, as opposed to a mere *distinctio rationis*. We are able to draw a conceptual distinction between the substance as mere substratum and the "full-blown" substance with qualities. Yet this distinction does not entail any real, metaphysical distinction. A substance is not a mereological complex entity consisting of simple entities such as a substratum, properties and genuine relations of support between substratum and properties. Rather, a property is a mode of the substance, one of its ways of being. Thus, if we are to drive a wedge between a substance's inherent causal impotence and its having powerful properties, we must assume an ontologically debatable separation between a substance and its properties.

Here my argument that agent causation is a serious alternative to event causation comes to an end. It should have become clear that, once a metaphysical framework of substances disposing of causal powers is established, agent causation fits naturally into it. In the final section I discuss evidence from developmental psychology and cognitive science indicating that the concept of agent causation is not only embedded in a particular metaphysical framework but is also deeply ingrained in our pre-theoretical grasp of ourselves and the world we inhabit.

3 Natural born agents?

As already indicated, the concept of agent causation is generally connected with free and rational actions.[13] The agent, so an oft-told story goes, has the unique

12 For this objection see Clarke (2003), pp. 188–193.

13 See for instance O'Connor (2001).

power to respond to reasons and to form intentions for actions accordingly. Thus agent causation is essentially intentional and purposive, in contrast to the blind processes of nature which can be reconstructed in event-causal terms. I have argued that there are metaphysical reasons for overcoming this dichotomy, because all real causation consists in substances acting upon each other. In this section I argue that there is epistemic evidence that, from the very beginning, our conceptual system is permeated by the idea of an agent being able to move its body spontaneously. If these epistemic data are true, they correspond with a metaphysics of powerful agents in broad terms and undermine the view to consider only intentional action as agent-caused.

3.1 Agent causation and developmental research

The ability to ascribe mental states such as beliefs, desires and intentions to other people begins rather late in child development and takes years to become fully functional. However, the ability to distinguish between self-moving goal-directed agents and entities in need of an external source of movement emerges much earlier.[14] Research suggests that 6-month-old infants already have a rudimentary capacity to distinguish between humans and inanimate objects in terms of goal-directed movements (Kuhlmeier et al. (2004)).

Spelke, Phillips and Woodward, for instance, discuss a study indicating that infants at this age do not apply what they call the "principle of contact" to the movement of human beings. (Spelke et al. (1995)). This principle says that physical objects move when another object comes into contact with them. In the study, 7-month-old infants were confronted with two different videotaped scenarios, one involving objects and another involving people. In the object-scenario, one inanimate object moved behind a screen and another emerged from the side of the screen. Infants looked longer at this scenario if the second object had begun to move before touching the first object. In the person-scenario, by contrast, a person moved behind the screen and a second person emerged from the side of the screen. If the second person had begun to move before coming into contact with the first, infants showed no signs of more attentively observing this scenario. Rather, they looked longer if the two people made contact first before the second person moved. These findings suggest that 7-month-old infants perceive only people, and not inanimate objects, as being capable of self-propulsion. Other studies complement these findings by showing that 9-month-olds consider the

14 This paragraph echoes mainly Steward (2009).

self-propulsion of an inanimate object like a robot anomalous, leading to negative reactions and emotional distress (Poulin-Dubois et al. (1996)).

This intriguingly secure grasp of spontaneous human movement in contrast to the motion of inanimate objects is not confined specifically to human agency. Gelman, for example, argues that humans are born with skeletal causal principles which, in combination with perceptual and other cues, lead us to acquire knowledge about animated and inanimated entities in general early on.[15]

She calls one principle the "innards principle". It says that self-propelled agents have insides that enable them to move on their own; she calls another the "external-agent principle", and this applies to entities that are not in a position to move on their own.

Moreover, infants around the age of one year seem to have a sophisticated non-mentalistic understanding of goal-directed actions which Gergely and Csibra call the "teleological stance".[16] According to the authors, infants at this age interpret actions as means to an end and evaluate the actions in the light of their efficacy. They can also generate inferences to identify relevant aspects of the action-context which justify the means even if the circumstances are not directly visible to them. The important point for our discussion is that the teleological stance does not involve a conscious ascription of mental states to the agents involved. Rather, it arises from the relationships among three elements: the action, the possible goal and the situational context. Once two of these three elements are given, 12-month-olds are capable of making an inference to the missing element by applying what the authors call the "rationality principle". This principle assumes that the agent will use the most effective means available in the situational context as the infant perceives it.

Without going into further details, the overall picture suggests that ample evidence supports the view that a basic conception of goal-directed and purposive agency, in contrast to an inanimate object's mechanistic movement, is part and parcel of our foundational conceptual make-up. One immediate consequence of this distinction appears to be that we directly conceive of animals moving their bodies, presuming that they possess a body which they move in a non-mechanistic way. We do not apply the principle of contact or the external agent principle in order to understand how an animal moves its body. Rather, we apply these principles to inanimate entities: we are inclined to say that there is a

15 Gelman (1990). See also Setoh et al. (2013) which confirms the assumption that the innards principle goes hand in hand with the ascription of basic biological properties, for instance that even quite diverse animals are not hollow.

16 Gergely and Csibra (2003).

certain part of this entity, the motor, whose function it is to set the entire object in motion mechanistically. Two further insights accompany this rather "holistic" understanding of an animal[17] and its body. On the one hand, one might say that, once an animal is said to possess a body, then that animal will in some sense be aware of possessing it. The distinction between the animal itself and its body seems to assume that the former has a certain subjective perspective on its body and the world which enables this distinction in the first place. If the animal lacked such a perspective, then one might wonder what supports the claim that the animal is not identical with its body. On the other hand, one might say that, once an animal is said to possess a body, then the animal will also exercise some form of control upon it. And controlling one's body — even if only minimally — is what grounds the capacity which humans experience as free will. Helen Steward makes this suggestion:

> Our natural inclination is to think of an animal as a creature that can, within limits, direct its own activities and which has certain choices about the details of those activities. (Steward (2009), p. 226)

This view converges nicely with the concept of agent-causation which claims, in a nutshell, that the agent has the capacity to bring about her activities directly. Agent-causation, then, not only has a natural home within a general metaphysical framework of powerful substances, but is also ubiquitous among animals. This is not to deny that some entities which we categorize as animals may ultimately, because their movements (contrary to appearances) are reducible to mere stimulus-reaction-mechanisms, turn out not to be true agents. Nor is it to conflate self-reflective rational agency with less complex forms of animal agency. The point is simply that there is no need to restrict agent causation to being the explanans of very special phenomena such as instances of full-blown rational decision-making.

3.2 Agent causation and enactivism

There is further evidence that agency should be considered to be a widespread and basic feature of our existence. Traditionally perception was seen as a passive process, in the sense that sensory input from the world enters the visual system

17 A traditional substance dualist picture of animated beings is opposed to this understanding of animal movements for it subscribes to a rather mechanistic understanding of how the mind moves the body. The latter is considered as a physical object related to the mind only externally.

and is converted in the brain into a mental image which is, ideally, a correct representation of the perceived object. The perceiver is conceived of, as Alva Noë has put it, as an automatic brain-photoreceptor system whose contents are static snapshot-like retinal images (Noë (2004)). For some time now this conception has been challenged by an alternative picture, so-called enactivism. Simply put, enactivism argues that perception is not a process in the brain whereby the perceptual apparatus constructs a mental representation out of the sensory input provided. Rather, the animal is actively engaged in perceiving, because perceiving itself is a skillful activity performed by the entire animal. This claim is based on the thesis that our perceptual apparatus is essentially connected with our sensorimotor and proprioceptive systems.

To illustrate this point, take vision as a paradigm model of perception. We tend to consider vision as a kind of photographic system: you open your eyes and, thanks to a complex internal process, a focused image of the world in front of you follows immediately. If movement is involved in this model at all, then it is merely as a means of adjusting your perspective in order to gain a better hold of what you wish more sharply to focus on. Moving the camera to the right place and taking the picture are two different events, related only externally. However, there is empirical evidence indicating that this picture is inadequate. Research about blindness, for instance, shows that there are forms of blindness that are not connected with dysfunctions in the visual system as such but rather with the organism's inability to integrate sensory input with patterns of movement. An example is given by attempts to restore vision to patients whom cataracts have made congenitally blind.[18] A cataract is a clouding of the lens of the eye which, in turn, affects vision. If the above camera-model were correct, then removing the cataract would result in removing the thing which impairs the animal's vision. Once the lens is cleared, light passes through to the retina unhindered, which should result in the animal's receiving a sharp image. Interestingly, however, case studies suggest that this does not happen. The surgery restores visual sensation, but this does not automatically restore the ability to see clearly. Immediately after the surgery, some patients continue to suffer a form of blindness. They report that their visual sensations are chaotic, confusing and uninformative to them.

From an enactivist perspective, the plausible explanation is that these patients cannot see because their visual impressions are not coupled with sensorimotor (self-)knowledge. In normal perceivers, sensation goes hand in hand with capacities for movement; we naturally turn our eyes towards an object of interest, we reach towards an object that catches our attention, we reflexively block our

18 The example is taken from Noë (2004), p. 4.

face with our hands if an object moves towards us. In all of these examples, sensory impressions are automatically coupled with spontaneous movement. One might say that the perceiving subject's visual impressions naturally fit the perceiver's movements because there is an implicit understanding that what is seen depends also on one's own body-posture and movements. The abovementioned patients seem to lack this sort of understanding. They fail to integrate the perceptual objects with their own changing movements (or the ways in which they might move over time), and this failure results in visual impressions which lack any useful content for the perceiver – who experiences this as a form of blindness.

There is further evidence that normal vision itself depends on self-produced movement and concurrent visual feedback. Held and Hein (1963) performed a classical study in which two kittens, one "active" and one "passive" were attached to an apparatus functioning like a carousel with black, white and metal-colored strips on the walls inside. The carousel was moved by the movements of the active kitten who was attached firmly but flexibly to it. The passive kitten was also attached to the apparatus but it was carried in a gondula. It could not move by itself but it was moved in the gondola by the movements of the other kitten. The apparatus was constructed in such a way that the gondula moved in accordance with the movements of the active kitten. The kittens could see neither each other nor their own limbs, but they were able to move their heads freely. Both kittens thus received the same visual input, but only the active kitten, because of its self-movement, received direct sensorimotor stimulation as well. The findings of this experiment are telling: only the active kittens developed normal depth-perception and visually guided paw placement responses. It seems that only through self-movement and concurrent visual feedback can animals develop functioning visually guided behavior. A foundational feature of perception is an implicit practical knowledge of how movements of one's body give rise to changes in sensory stimulation.

If we adopt an enactivist standpoint on vision, then we might not only question the assumption that vision amounts to a passive process of internally representing the world, but we might go a step farther. If an animal is essentially an active embodied being situated in a determinate environment, then why assume that an internal representation intervening between the animal and the world is needed at all? Why not simply suppose that the world is immediately present to the animal? If you want to reach out for a cup of tea, then why assume that doing so requires an internal representation of the cup in front of you? The alternative enactivist account suggests that the very directing of your gaze to the cup amounts to a direct perceiving of the cup as something reachable. The idea is that the cup assumes the role of guiding the hand in your act of reaching for it. In other words, representation may not be required in order for action to

follow on. If animals have fundamentally agential natures, then they may perceive the world directly as full of opportunities for action. This echoes Heidegger's analysis of being-in-the-world which emphasizes that our primary understanding of the world is not one of objects describable in terms of numbers, measures and weights, but of a world loaded with references for use. We perceive the world primarily from the perspective of agents, not observers.

These considerations should suffice to motivate the claim that an enactivist model of perception can support the concept of agent causation proposed in this article. If the agential nature fundamentally shapes the animal's being in the world, then agent-causal terms may provide the most adequate metaphysical reconstruction of this nature. The animal's experience of the world presupposes its active engagement with the world. Any reconstruction in non-agential terms seems to miss the most basic features of what it means for an animal to be alive. This view nicely complements Steward's metaphysical conception of the animal as a self-moving entity which executes some form of direct control over its body. And it might also help us explain why the distinction between animated and inanimated beings figures so centrally in our conceptual scheme. Being self-moving animals ourselves, it is unsurprising that this basic existential feature is mapped into our basic understanding of the world.

4 Conclusion

The contemporary discussion of agent causation focuses on the causal production of free rational action, where such action is seen in radical opposition to the omnipresent event-causal processes which determine natural phenomena. However, if the above arguments are correct, then all causation, whether animate or inanimate, can be modelled along the structural features of agent-causation. First, causation always involves one substance acting on, or being acted on by, another substance. Second, within the animal kingdom agent causation is a mundane phenomenon, because animals themselves are natural-born agents. Third, as a consequence, the production of free rational action is a variation of ordinary animal agent causation brought about by rational animals. One might wonder about the bad philosophical press which agent causation thus far received. Perhaps this reputation has less to do with the concept of agent causation itself than with a long concatenation of philosophical distortions and biases. Substance-dualist worries, empiricist epistemological meticulousness, an overemphasis on mechanistic-reductionist thinking, and a deep mistrust of our commonsense reasoning may have obscured our view of something right under our noses: the fact that, when we experience and interact with the world, we are first and foremost agents.

Alessandro Giordani

Quantified Modal Justification Logic with Existence Predicate

1 Introduction

Systems of explicit modal logic provide a powerful framework for characterizing modal concepts as arising from the existence of appropriate sources. It is well known that, within the standard semantic framework of epistemic logic, the notion of knowledge is typically defined by assuming that a proposition is known to be true precisely when it is true in every epistemically possible world. If we adopt the view that evidence is the source of knowledge, we are led to endorse an analysis of knowledge according to which a poposition is known to be true precisely when it is accepted as true on the basis of some piece of evidence, so that the truth of that proposition in every epistemically possible world can be viewed as grounded on the existence of some piece of evidence for it[1]. Similarly, within the standard semantic framework of modal logic the notion of necessary truth is characterized by assuming that a proposition is necessarily true precisely when it is true in every ontically possible world. So, if we adopt the view that essence is the source of necessity, we are led to endorse an analysis of necessity according to which a poposition is necessarily true precisely when it is true on the basis of the essence of the entities to which it is referring, so that the truth of that proposition in every ontically possible world can be viewed as grounded on the existence of essences[2]. In what follows, we will develop these intuitions by constructing a system of quantified modal logic of justification, endowed with a specific predicate of existence, which can help us to model both situations where the distinction between actual and potential possession of a justification is crucial and situations

This paper is dedicated to Sergio Galvan, who first introduced me to modal logic and metamathematics. He inspired and encouraged me in my researches during the years and is an admirable example of philosophical insight and passion for the truth.

1 Systems of explicit epistemic logic are known as justification logics, which are development of systems of logic of proofs originally proposed by Artemov. See Artemov (2001, 2008); Artemov and Nogina (2005); Fitting (2005) for a general introduction.

2 Systems of explicit ontic logic are developed in Giordani (2013) for interpreting the logic of essence.

Alessandro Giordani: Università Cattolica, Milan

DOI: 10.1515/9783110529494-013

where sources of knowldge are subjected to change.[3] The paper is structured as follows. In section 2 the basic system of quantified modal justification logic we are interested in is introduced from an axiomatic point of view. In section 3 the system previously introduced is proved to be sound and complete with respect to a suitable possibile world semantics. Finally, the system is extended in order to incorporate a predicate of existence, to be intended as indicating acknowledged existence, and some observations are made on its usefulness.

2 The main system

2.1 The language of justification logic

In this section the basic language $\mathcal{L}$ of the systems of quantified modal logic of justification is proposed.

Definition 1. *Terms of $\mathcal{L}$.*
The set $Tm(\mathcal{L})$ of terms of $\mathcal{L}$ is defined according to the following rules:

$$t := c \mid x \mid (t \cdot t) \mid (t + t) \mid !t \mid gen_x(t)$$

where x is an element of a countable set $\{x_i\}_{i\in\mathbb{N}}$ of variables and c is an element of a countable set $\{c_i\}_{i\in\mathbb{N}}$ of constants.

Definition 2. *Formulas of $\mathcal{L}$.*
The set $Fm(\mathcal{L})$ of formulas of $\mathcal{L}$ is defined according to the following rules:

$$\varphi := p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid \forall x\varphi \mid \Box\varphi \mid t : \varphi$$

where p is an element of a countable set $\{p_i\}_{i\in\mathbb{N}}$ of propositional variables and t is a term in $Tm(\mathcal{L})$.

In what follows, composite terms are said to be *closed* if no free variable occurs in them, where all the variables constituting a term are counted as free except the ones occurring as indices in terms like $gen_x(t)$.

Intuitively, terms in $Tm(\mathcal{L})$ are interpreted on structured sources of justification and formulas like $t : \varphi$ are interpreted as stating that t is a source of

3 A reference possible world semantics for quantified justification logic is proposed in Fitting (2008).

justification for φ. It is then assumed that sources of justification, or *justifiers*, are abstract entities provided with a specific structure and subjected to a set of operations. In particular, the operations $\cdot$, $+$, $!$ are used to construct new sources from basic ones. Here, $t \cdot s$ is intended as a justifier that provides justication for all the sentences that can be justified by applying *modus ponens* to premises justified by t and by s, while $t + s$ is intended as a justifier providing justication for all the sentences that can be justified either by t or by s. In addition, $!$ is a justication checker: it returns a justifier $!t$ for the sentence stating that t is a source of justification for φ, provided that t is indeed such a source.

2.2 Axiomatization

Definition 3. *Axioms of quantified modal justification logic.*

Let us introduce the following groups of axioms[4].

Group I: propositional axioms and *modus ponens*.

Any set of classical propositional axioms is appropriate

Group II: quantification axioms

Q1: $\forall x(\varphi_1 \to \varphi_2) \to (\varphi_1 \to \forall x\varphi_2)$, x not free in φ

Q2: $\forall x\varphi \to \varphi^t_x$, t free for x in φ

RQ: if $\vdash \varphi$, then $\vdash \forall x\varphi$

Group III: modal axioms

N1: $\Box(\varphi_1 \to \varphi_2) \to (\Box\varphi_1 \to \Box\varphi_2)$

N2: $\forall x\Box\varphi \to \Box\forall x\varphi$

N3: $\Box\varphi \to \varphi$

N4: $\Box\varphi \to \Box\Box\varphi$

N5: $\neg\Box\varphi \to \Box\neg\Box\varphi$

RN: if $\vdash \varphi$, then $\vdash \Box\varphi$

Group IV: justification axioms

J1: $t_1 : (\varphi_1 \to \varphi_2) \to (t_2 : \varphi_1 \to t_1 \cdot t_2 : \varphi_2)$

J2: $t_1 : \varphi \vee t_2 : \varphi \to t_1 + t_2 : \varphi$

J3: $t : \varphi \to !t : t : \varphi$

J4: $t : \varphi \to \varphi$

4 Axioms are introduced schematically. Accordingly, an axiom can be identified with the set of its instances. Note also that in group III axioms **N2** and **N4** can be derived from the other ones plus the axioms in group II.

J5: $t : \varphi \rightarrow \Box\varphi$, where t is closed
JQ: $\forall x(t : \varphi) \rightarrow gen_x(t) : \forall x\varphi$
RJ: if $c : \varphi \in \mathcal{C}$, then $\vdash c : \varphi$, where $\mathcal{C}$ is a constant specification

A *constant specification* for $\mathcal{L}$ is any set of formulas $c : \varphi$, where φ is formula of $\mathcal{L}$ and c is a constant. The general idea is that constants are sources for the justification of formulas and constant specifications are relations connecting sources and justified formulas. In the present context, a constant specification is said to be *axiomatically appropriate*, if all the axioms are justified by some constant, and uniform, if $c : \varphi \in \mathcal{C} \Rightarrow c : \varphi_x^y \in \mathcal{C}$, for every x and every y that is free for x in φ. We will only work with axiomatically appropriate uniform constant specifications, so that any axioms turn out to be justified by some source and any source that justifies an axiom instance also justifies all of its variants.

Definition 4. *System* $\mathbf{QLJ_{S5}}$.

A system of quantified modal justification logic $\mathbf{QLJ}(\mathcal{A}, \mathcal{C})$ is characterized by axioms of groups I, II, III, a particular selection $\mathcal{A}$ of axioms from group IV, and a particular axiomatically appropriate uniform constant specification $\mathcal{C}$. The system we are primarily interested in is

– $\mathbf{QLJ_{S5}} = \mathbf{QLJ}(\mathcal{C}, \mathcal{A})$, with $\mathcal{A}$ = group IV

System $\mathbf{QLJ_{S5}}$ is obtained by combining, using **JQ** and **J5**, the basic system **LJ** of logic of factive justification, given by the axioms of group I, plus **J1**, **J2**, **J3**, **J4**, and **RJ**, with a standard **QS5** system of quantified logic for ontic necessity, which in turn is obtained by combining the standard **S5** system of propositional modal logic for ontic necessity with a system of first order logic, as axiomatized by the axioms and rules of groups I-II. In addition, in this system $\mathcal{C}$ is required to be such that $c : \varphi \in \mathcal{C}$ only if φ is an axiom of groups I-IV. This last condition ensures that only logical axioms and and logical theorems are justified by closed terms.

Theorem 1. *if* $\vdash_{\mathbf{QLJ_{S5}}} \varphi$, *then* $\vdash_{\mathbf{QLJ_{S5}}} t : \varphi$, *for some closed term* t.

Proof. By induction on the length of a derivation. If φ is an axiom instance, then the conclusion follows from **RJ**, since all the constants are closed terms. If φ is derived from $\vdash \varphi$ and $\vdash \varphi \rightarrow \psi$ by *modus ponens*, then $\vdash s : \varphi$ and $\vdash t : (\varphi \rightarrow \psi)$, where both s and t are closed, by induction hypothesis, and the conclusion follows by **J1**, given that $t \cdot s$ is closed. Suppose φ is derived by **RJ**. Then φ has the form $c : \psi$ for some ψ. Hence, $\vdash !c : c : \psi$, by **J4**, with $!c$ closed. Suppose φ is

derived by **RQ**. Then φ has the form $\forall x\psi$ for some ψ, and $\vdash \psi$. Hence, by induction hypothesis, $\vdash t : \psi$, with t closed, and so $\vdash \forall x(t : \psi)$ by **Q3**. Thus, $\vdash gen_x(t) : \forall x\psi$, by **J2**, and the conclusion follows, since x is not free in $gen_x(t)$. Suppose, finally, that φ is derived by **RN**. Then φ has the form $\Box\psi$ for some ψ, and $\vdash \psi$. Thus, by induction hypothesis, $\vdash t : \psi$, with t closed. Since $\vdash t : \psi \rightarrow \Box\psi$, by **JN**, and $\vdash !t : t : \psi$, by **J4**, we obtain $\vdash c : (t : \psi \rightarrow \Box\psi)$, by **RJ**, so that $\vdash c \cdot !t : \Box\psi$, by **J1**, and the conclusion follows, given that $c \cdot !t$ closed. □

Hence, $\mathbf{QLJ_{S5}}$ enjoys internalization, since every derivable formula can be associated with a piece of evidence justifying it.

3 Semantics

The semantics for $\mathcal{L}$ is based on the notions of justification model and modal model, and constitutes an extension of the semantics for the logic of justification. A justification model provides references for the terms in $Tm(\mathcal{L})$, while a modal model allows us to define the notion of truth with respect to the formulas in $Fm(\mathcal{L})$. The fundamental idea underlying the semantics of the logic of justification is that an agent is justified in assuming that a certain proposition is true precisely when that proposition holds in all the epistemic worlds that are admitted by the agent, i.e., in all the epistemic worlds that are compatible with her epistemic state. Furthermore, the agent has a justifier for assuming that a certain proposition is true precisely when she is justified in believing the proposition by virtue of possessing a justifier which is admissible for it. Finally, it is possible for the agent to acquire a justifier for assuming that a certain proposition is true precisely when there is a possible world where she actually has a justifier for assuming it. As a consequence, in order to model the epistemic state of an agent, we have to introduce the following elements: (1) a set of epistemic worlds; (2) a relation of ontic accessibility between them, determining which worlds are possible with respect to any world; (3) a relation of epistemic accessibility between them, determining which worlds are compatible with the epistemic state of the agent; (4) a set of justifiers, provided with a set of operations allowing us to construct composite justifiers from elementary ones; (5) a selection function, determining which justifiers are admissible for each proposition at each epistemic world.

Let us now present all the elements in turn.

3.1 Justification model

Definition 5. *justification frame for $\mathcal{L}$.*

A justification frame for $\mathcal{L}$ is a tuple $\mathfrak{D} = \langle D, \cdot^D, +^D, !^D, gen^D \rangle$ where

(i) $D \neq \varnothing$
(ii) $\cdot^D : D \times D \to D$
(iii) $+^D : D \times D \to D$
(iv) $!^D : D \to D$
(v) $gen^D : (D \to D) \to D$

Intuitively, D is a set of justifiers and, in accordance with the intended interpretation of the language anticipated in the first section: $\cdot^D$ is a function taking pairs $\langle d_1, d_2 \rangle$ of justifiers and returnig a justifier for every proposition that is the consequent of an implication justified by d_1 and an antecedent justified by d_2; $+^D$ is a function taking pairs $\langle d_1, d_2 \rangle$ of justifiers and returnig a justifier for every proposition justified by either d_1 or d_2; $!^D$ is a function taking a justifier d and returnig a justifier checking whether d justifies a proposition; finally, gen^D is a function taking functions from justifiers to justifiers and returnig a new kind of justifier, whose characterization is presented below.

Definition 6. *justification model for $\mathcal{L}$.*

A justification model for $\mathcal{L}$ is a tuple $\langle \mathfrak{D}, \mathfrak{I} \rangle$ where $\mathfrak{D}$ is a justification structure and $\mathfrak{I} : Tm(\mathcal{L}) \to D$ is an interpretation mapping both variables and constants of $\mathcal{L}$ on elements of D. As usual $\mathfrak{I}_x^d$ is the re-interpretaion of x on d based on $\mathfrak{I}$, which is the interpretation such that

$$\mathfrak{I}_x^d(y) = \begin{cases} d & \text{if } x = y \\ \mathfrak{I}(x) & \text{otherwise} \end{cases}$$

$\mathfrak{I}$ can be extended to a valuation $\mathfrak{I}[\,]$ of all terms in $Tm(\mathcal{L})$ by setting

$\mathfrak{I}[x] = \mathfrak{I}(x)$

$\mathfrak{I}[c] = \mathfrak{I}(c)$

$\mathfrak{I}[t_1 \cdot t_2] = \mathfrak{I}[t_1] \cdot^D \mathfrak{I}[t_2]$

$\mathfrak{I}[t_1 + t_2] = \mathfrak{I}[t_1] +^D \mathfrak{I}[t_2]$

$\mathfrak{I}[!t] = !^D \mathfrak{I}[t]$

$\mathfrak{I}[gen_x(t)] = gen^D(\lambda \mathfrak{I}(x).\mathfrak{I}[t])$

In the last condition $\lambda\mathfrak{I}(x).\mathfrak{I}[t]$ is the function such that $\lambda\mathfrak{I}(x).\mathfrak{I}[t](d) = \mathfrak{I}_x^d[t]$, based on the fact that $\mathfrak{I}[t] = \mathfrak{I}_x^{\mathfrak{I}(x)}[t]$.

3.2 Modal model

Definition 7. *modal frame for* $\mathcal{L}$.

A modal frame for **QLJ$_{S5}$** *is a tuple* $\langle W, R, K, \mathfrak{D}, \mathfrak{I}, E\rangle$ *where*

- W *is a set of worlds*
- $R \subseteq W \times W$ *is an equivalence relation of ontic accessibility*
- $K \subseteq W \times W$ *is a reflexive and transitive relation of epistemic accessibility*
- $\langle \mathfrak{D}, \mathfrak{I}\rangle$ *is a justification model*
- $E : W \times D \times Int_D \to \wp(Fm(\mathcal{L}))$ *is a selection function assigning to each triple* $(w, d, \mathfrak{I})$ *of worlds, justifiers, and interpretations the sets of formulas of* $\mathcal{L}$ *justified by* d *at* w *under interpretation* $\mathfrak{I}$.

E is required to satisfy the following closure conditions, where $E(w, d_1, \mathfrak{I}) \circ E(w, d_2, \mathfrak{I})$ is the set containing the consequents of implications contained in $E(w, d_1, \mathfrak{I})$ with antecedents in $E(w, d_2, \mathfrak{I})$.

E1: $E(w, d_1, \mathfrak{I}) \circ E(w, d_2, \mathfrak{I}) \subseteq E(w, d_1 \cdot^D d_2, \mathfrak{I})$

E2: $E(w, d_1, \mathfrak{I}) \cup E(w, d_2, \mathfrak{I}) \subseteq E(w, d_1 +^D d_2, \mathfrak{I})$

E3: $\varphi \in E(w, d, \mathfrak{I})$ and $d = \mathfrak{I}(t) \Rightarrow t : \varphi \in E(w, !^D d, \mathfrak{I})$

E4: $\forall d \in D(\varphi \in E(w, \mathfrak{I}_x^d[t], \mathfrak{I}_x^d)) \Rightarrow \forall x\varphi \in E(w, \mathfrak{I}[gen_x(t)], \mathfrak{I})$

E5: $\varphi \in E(w, \mathfrak{I}[t], \mathfrak{I}) \Rightarrow \forall v(R(w, v) \Rightarrow \mathcal{M}, v \vDash \varphi)$, if t is closed

E6: $\{\varphi \mid c : \varphi \in \mathcal{C}\} \subseteq E(w, \mathfrak{I}(c), \mathfrak{I})$

E7: $\mathfrak{I}_1 \overset{\varphi}{=} \mathfrak{I}_2 \Rightarrow \varphi \in E(w, d, \mathfrak{I}_1) \Leftrightarrow \varphi \in E(w, d, \mathfrak{I}_2)$

EM: $K(w, v) \Rightarrow E(w, d, \mathfrak{I}) \subseteq E(v, d, \mathfrak{I})$

where, in *E6*, $\overset{\varphi}{=}$ is identity relative to the values assigned to free variables in φ. *EM* is a monotonicity requirement, which states that all the evidence possessed in a world w is preserved at any world to which w has an access.

Definition 8. *modal model for* **QLJ$_{S5}$**.

A modal model for $\mathcal{L}$ *is a tuple* $\mathcal{M} = \langle W, R, K, \mathfrak{D}, \mathfrak{I}, E, V\rangle$ *where*

(i) $\langle W, R, K, \mathfrak{D}, \mathfrak{I}, E\rangle$ *is a modal frame for* $\mathcal{L}$ *and*

(ii) $V : W \times Var \to \{0, 1\}$ *is a valuation of the propositional variables of* $\mathcal{L}$.

If $\mathcal{M} = \langle W, R, K, \mathfrak{D}, \mathfrak{I}, E, V\rangle$, then $\mathcal{M}_x^d = \left\langle W, R, K, \mathfrak{D}, \mathfrak{I}_x^d, E, V\right\rangle$, where $\mathfrak{I}_x^d$ is the re-interpretaion of x on d based on $\mathfrak{I}$.

Definition 9. *truth at a world in a model for* $\mathbf{QLJ_{S5}}$.

$\mathcal{M}, w \vDash p \Leftrightarrow V(w,p) = 1$

$\mathcal{M}, w \vDash \neg\varphi \Leftrightarrow \mathcal{M} \nvDash \varphi$

$\mathcal{M}, w \vDash (\varphi_1 \wedge \varphi_2) \Leftrightarrow \mathcal{M} \vDash \varphi_1$ *and* $\mathcal{M} \vDash \varphi_2$

$\mathcal{M}, w \vDash \forall x\varphi \Leftrightarrow \forall d \in D(\mathcal{M}_x^d, w \vDash \varphi)$

$\mathcal{M}, w \vDash \Box\varphi \Leftrightarrow \forall v(R(w,v) \Rightarrow \mathcal{M}, v \vDash \varphi)$

$\mathcal{M}, w \vDash t : \varphi \Leftrightarrow \forall v(K(w,v) \Rightarrow \mathcal{M}, v \vDash \varphi)$ *and* $\varphi \in E(w, \mathfrak{I}[t], \mathfrak{I})$

Corollary 1. (*Monotonicity*). $\mathcal{M}, w \vDash t : \varphi \Rightarrow \forall v(K(w,v) \Rightarrow \mathcal{M}, v \vDash t : \varphi)$.

Proof. Suppose $\mathcal{M}, w \vDash t : \varphi$. Then $\forall v(K(w,x) \Rightarrow \mathcal{M}, x \vDash \varphi)$ and $\varphi \in E(w, \mathfrak{I}[t], \mathfrak{I})$. Suppose $K(w,v)$. Then $\forall x(K(v,x) \Rightarrow K(w,x))$, since K is transitive, and so $\forall x(K(v,x) \Rightarrow \mathcal{M}, x \vDash \varphi)$, since $K(w,x) \Rightarrow \mathcal{M}, x \vDash \varphi$. Thus, $K(w,v) \Rightarrow \forall x(K(v,x) \Rightarrow \mathcal{M}, x \vDash \varphi)$ and $K(w,v) \Rightarrow \varphi \in E(v,d,\mathfrak{I})$, by *EM*, and so $\forall v(K(w,v) \Rightarrow \mathcal{M}, v \vDash t : \varphi)$. □

The definition of the relation of logical consequence is the usual one: $X \Vdash_{\mathbf{QLJ_{S5}}} \varphi$ iff $M, w \vDash X \Rightarrow M, w \vDash \varphi$, for every M, w, where $M, w \vDash X$ precisely when $M, w \vDash \varphi$ for every $\varphi \in X$. In what follows, we will also write

$R(w)$	for $\{v \mid R(w,v)\}$
$K(w)$	for $\{v \mid K(w,v)\}$
$M, R(w) \vDash \varphi$	for $\forall v(R(w,v) \Rightarrow M, v \vDash \varphi)$
$M, K(w) \vDash \varphi$	for $\forall v(K(w,v) \Rightarrow M, v \vDash \varphi)$

In the next two sections we will prove our main result, which is that system $\mathbf{QLJ_{S5}}$ is sound and strongly complete with respect to the class of modal model for $\mathbf{QLJ_{S5}}$.

3.3 Soundness

$\mathbf{QLJ_{S5}}$ is sound with respect to the class of modal model for $\mathbf{QLJ_{S5}}$.

Theorem 2. (*Soundness*).
$X \Vdash_{\mathbf{QLJ_{S5}}} \varphi \Rightarrow X \vdash_{\mathbf{QLJ_{S5}}} \varphi$, *for all* X *and* φ.

We will only focus on the validity of the axioms concerning proofs. The proofs of the validity of axioms in groups I-III are standard.

J1: $t_1 : (\varphi_1 \to \varphi_2) \to (t_2 : \varphi_1 \to t_1 \cdot t_2 : \varphi_2)$

Proof. Suppose $\mathcal{M}, w \vDash t_1 : (\varphi_1 \to \varphi_2)$ and $\mathcal{M}, w \vDash t_2 : (\varphi_1)$. Then
$\forall v(K(w,v) \Rightarrow \mathcal{M}, v \vDash \varphi_1 \to \varphi_2)$ and $\varphi_1 \to \varphi_2 \in E(w, \mathfrak{J}[t_1], \mathfrak{J})$, by def. $\vDash$
$\forall v(K(w,v) \Rightarrow \mathcal{M}, v \vDash \varphi_1)$ and $\varphi_1 \in E(w, \mathfrak{J}[t_2], \mathfrak{J})$, by def. $\vDash$
$\forall v(K(w,v) \Rightarrow \mathcal{M}, v \vDash \varphi_2)$ and $\varphi_2 \in E(w, \mathfrak{J}[t_1 \cdot t_2], \mathfrak{J})$, by *E1* □

J2: $t_1 : \varphi \vee t_2 : \varphi \to t_1 + t_2 : \varphi$

Proof. Suppose $\mathcal{M}, w \vDash t_i : \varphi$, $i = 1, 2$. Then
$\forall v(K(w,v) \Rightarrow \mathcal{M}, v \vDash \varphi)$ and $\varphi \in E(w, \mathfrak{J}[t_i], \mathfrak{J})$, by def. $\vDash$
$\forall v(K(w,v) \Rightarrow \mathcal{M}, v \vDash \varphi)$ and $\varphi \in E(w, \mathfrak{J}[t_1 + t_2], \mathfrak{J})$, by *E2*
$\mathcal{M}, w \vDash t_1 + t_2 : \varphi$, by def. $\vDash$ □

J3: $t : \varphi \to !t : t : \varphi$

Proof. Suppose $\mathcal{M}, w \vDash t : \varphi$. Then
$\forall v(K(w,v) \Rightarrow \mathcal{M}, v \vDash \varphi)$ and $\varphi \in E(w, \mathfrak{J}[t], \mathfrak{J})$, by def. $\vDash$
$\forall v(K(w,v) \Rightarrow \mathcal{M}, v \vDash t : \varphi)$ and $t : \varphi \in E(w, !^D\mathfrak{J}[t], \mathfrak{J})$, by monotonicity and *E3*
$\mathcal{M}, w \vDash !t : t : \varphi$, by def. $\vDash$ □

J4: $t : \varphi \to \varphi$

Proof. Suppose $\mathcal{M}, w \vDash t : \varphi$. Then
$\forall v(K(w,v) \Rightarrow \mathcal{M}, v \vDash \varphi)$ and $\varphi \in E(w, \mathfrak{J}[t], \mathfrak{J})$, by def. $\vDash$
$\mathcal{M}, w \vDash \varphi$, since K is reflexive
$\mathcal{M}, w \vDash \varphi$, by def. $\vDash$ □

J5: $t : \varphi \to \Box\varphi$, t closed

Proof. Suppose $\mathcal{M}, w \vDash t : \varphi$. Then
$\forall v(K(w,v) \Rightarrow \mathcal{M}, v \vDash \varphi)$ and $\varphi \in E(w, \mathfrak{J}[t], \mathfrak{J})$, by def. $\vDash$
$\forall v(R(w,v) \Rightarrow \mathcal{M}, v \vDash \varphi)$, by *E5*, since t is closed
$\mathcal{M}, w \vDash \Box\varphi$, by def. $\vDash$ □

JQ: $\forall x(t : \varphi) \to gen_x(t) : \forall x\varphi$.

Proof. Suppose $\mathcal{M}, w \vDash \forall x(t : \varphi)$.

$\forall d \in D(\mathcal{M}^d_x, w \vDash t : \varphi)$, by def. $\vDash$
$\forall d \in D(\forall v(R(w,v) \Rightarrow \mathcal{M}^d_x, v \vDash \varphi)$ and $\varphi \in E(\mathfrak{I}^d_x[t], \mathfrak{I}^d_x))$, by def. $\vDash$
$\forall d \in D(\forall v(R(w,v) \Rightarrow \mathcal{M}^d_x, v \vDash \varphi))$ and $\forall d \in D(\varphi \in E(\mathfrak{I}^d_x[t], \mathfrak{I}^d_x))$, by logic
$\forall v(R(w,v) \Rightarrow \forall d \in D(\mathcal{M}^d_x, v \vDash \varphi))$ and $\forall d \in D(\varphi \in E(\mathfrak{I}^d_x[t], \mathfrak{I}^d_x))$, by logic
$\forall v(R(w,v) \Rightarrow \forall d \in D(\mathcal{M}^d_x, v \vDash \varphi))$ and $\forall x\varphi \in E(\mathfrak{I}[gen_x(t)], \mathfrak{I})$, by *E4*
$\forall v(R(w,v) \Rightarrow \mathcal{M}, v \vDash \forall x\varphi)$ and $\forall x\varphi \in E(\mathfrak{I}[gen_x(t)], \mathfrak{I})$, by def. $\vDash$
$\mathcal{M}, w \vDash gen_x(t) : \forall x\varphi$, by def. $\vDash$ □

RJ: if $c : \varphi \in \mathcal{C}$, then $\vdash c : \varphi$.

Straightforward, given *E6*.

3.4 Completeness

Let us now focus on the completeness theorem.

Theorem 3. *completeness.*
$X \Vdash_{\mathbf{QLJ_{ss}}} \varphi \Rightarrow X \vdash_{\mathbf{QLJ_{ss}}} \varphi$, *for all* X *and* φ.

The proof is by construction of a canonical model. As usual, we start with a consistent set of formulas X and extend it to a maximally consistent and witnessed set X^* in a language $\mathcal{L}^*$ containing a countable set $\{n_i\}_{i\in\mathbb{N}}$ of new constants, where a set X^* is said to be witnessed provided that, for every φ and every variable x, there is some variable y such that $\varphi^y_x \rightarrow \forall x\varphi \in X^*$. X^* is then such that

1. $X \subseteq X^*$
2. $\varphi \in X^* \Leftrightarrow \neg\varphi \notin X^*$
3. $(\varphi_1 \wedge \varphi_2) \in X^* \Leftrightarrow \varphi_1 \in X^*$ and $\varphi_2 \in X^*$
4. $\forall x\varphi \in X^* \Leftrightarrow \varphi^t_x \in X^*$ for all $t \in Tm(\mathcal{L}^*)$

We will also exploit the following well-known facts, where $w/\Box = \{\psi \mid \Box\psi \in w\}$:

Fact 1. *If* X *is a consistent subset of* $Fm(\mathcal{L})$ *then there is a consistent witnessed subset of* $Fm(\mathcal{L}^*)$ *which includes* X.

Fact 2. *If* X *is a maximally consistent witnessed subset of* $Fm(\mathcal{L}^*)$ *and* $\Box\varphi \notin X$, *then* $w/\Box, \neg\varphi$ *is included in a consistent witnessed subset of* $Fm(\mathcal{L}^*)$.

Proof. See Hughes and Cresswell (1996) (theorems 14.1 and 14.2, pp. 258–261). Axiom $\forall x\Box\varphi \rightarrow \Box\forall x\varphi$ is essential for proving fact 2. □

The constant specification $\mathcal{C}$ has to be extended accordingly. In particular, a new constant specification $\mathcal{C}^*$ is defined from $\mathcal{C}$ as follows. Let $\varphi \in \mathcal{L}^*$ be any new axiom instance and $n_1 \dots n_k$ be the sequence of new constants occuring in it. Then $\varphi \in \mathcal{C}^*(c)$ precisely when there is a corresponding instance $\varphi^{x_1 \dots x_k}_{n_1 \dots n_k} \in \mathcal{C}(c)$, where $x_1 \dots x_k$ is a sequence of old variables not occurring in φ. Since φ is an axiom instance, $\varphi^{x_1 \dots x_k}_{n_1 \dots n_k}$ is an axiom instance, so that $\varphi^{x_1 \dots x_k}_{n_1 \dots n_k} \in \mathcal{C}(c)$ for some constant c. Since $\mathcal{C}$ is axiomatically appropriate and uniform, $\mathcal{C}^*$ is axiomatically appropriate and uniform by definition. Coming back to the construction of the canonical model, the more involved part concerns the construction of a suitable justification frame for the new language $\mathcal{L}^*$.

Definition 10. *canonical justification frame for $\mathcal{L}^*$.*
The canonical justification frame for $\mathcal{L}^*$ is a tuple $\langle \mathfrak{D}, \mathfrak{I} \rangle$ where

- $\mathfrak{D}$ is defined as follows:
  - D is the set of terms of $\mathcal{L}^*$
  - $\cdot^D$ is such that $d_1 \cdot^D d_2 = (d_1 \cdot d_2)$
  - $+^D$ is such that $d_1 +^D d_2 = (d_1 + d_2)$
  - $!^D$ is such that $!^D d = !d$
  - gen^D is such that $gen^D(\lambda x.t) = gen_x(t)$

  where $\lambda x.t$ is the function such that $\lambda x.t(d) = t^d_x$.
- $\mathfrak{I}$ is the canonical assignment such that $\mathfrak{I}(t) = t$, for all terms t.

Lemma 1. *Let t be a term and y a variable in t. Then $\mathfrak{I}^d_y[t] = t^d_y$.*

Proof. by induction on the length of a term.

Case 1: $y = x$. Then $\mathfrak{I}^d_y[x] = d = y^d_y$

Case 2: $y \neq x$. Then $\mathfrak{I}^d_y[x] = x = x^d_y$

$\mathfrak{I}^d_y[c] = \mathfrak{I}^d_y(c) = c = c^d_y$

$\mathfrak{I}^d_y[t_1 \cdot t_2] = \mathfrak{I}^d_y[t_1] \cdot^D \mathfrak{I}^d_y[t_2] = t_1{}^d_y \cdot t_2{}^d_y = (t_1 \cdot t_2)^d_y$

$\mathfrak{I}^d_y[t_1 + t_2] = \mathfrak{I}^d_y[t_1] +^D \mathfrak{I}^d_y[t_2] = t_1{}^d_y + t_2{}^d_y = (t_1 + t_2)^d_y$

$\mathfrak{I}^d_y[!t] = !^D \mathfrak{I}^d_y[t] = !t^d_y = (!t)^d_y$

As to $gen_x(t)$, note that $\lambda \mathfrak{I}(x).\mathfrak{I}[t](d) = \mathfrak{I}^d_x[t] = t^d_x$, by the inductive hypothesis, and that $\mathfrak{I}^d_y[gen_x(t)] = \mathfrak{I}[gen_x(t)]$, since x is not free in $gen_x(t)$.

Case 1: $y = x$. Then $\mathfrak{I}^d_y[gen_x(t)] = gen^D(\lambda \mathfrak{I}(x).\mathfrak{I}[t])$

Thus $\mathfrak{J}_y^d[gen_x(t)] = gen^D(\lambda x.t) = gen_x(t) = (gen_x(t))_y^d$

Case 2: $y \neq x$. Then $\mathfrak{J}_y^d[gen_x(t)] = gen^D(\lambda \mathfrak{J}_y^d(x).\mathfrak{J}_y^d[t])$

Thus $\mathfrak{J}_y^d[gen_x(t)] = gen^D(\lambda x.t_y^d) = gen_x(t_y^d) = (gen_x(t))_y^d$ □

Corollary 2. *Let φ be a formula. Then $M_x^d, w \vDash \varphi \Leftrightarrow M, w \vDash \varphi_x^d$.*

Proof. Straightforward, by the previous lemma. □

Definition 11. *canonical model*

The canonical modal model for $\mathcal{L}^*$ is the tuple $\langle W, R, K, \mathfrak{D}, \mathfrak{J}, E, V\rangle$ where

- *$\langle \mathfrak{D}, \mathfrak{J}\rangle$ is the canonical justification frame*
- *W is the set of maximally consistent sets in $\mathcal{L}^*$*
- *R is such that $R(w,v) \Leftrightarrow w/\Box = \{\varphi \mid \Box\varphi \in w\} \subseteq v$*
- *K is such that $K(w,v) \Leftrightarrow w/K = \{\varphi \mid \exists t(t : \varphi \in w)\} \subseteq v$*
- *E is such that $E(w,d,\mathfrak{J}_x^{d'}) := \{\varphi \mid \varphi_x^{d'} \in w/d\}$*
- *V is such that $V(p) = \{w \mid p \in w\}$*

Hence, the set of formulas justified by d under $\mathfrak{J}_x^{d'}$ is the set of variants of formulas in w/d where d' is substituted for x.

Lemma 2. *(Canonicity Lemma). Let $M = \langle W, R, K, \mathfrak{D}, \mathfrak{J}, E, V\rangle$ be the canonical model for* $\mathbf{QLJ_{S5}}$. *Then M is a model for* $\mathbf{QLJ_{S5}}$.

Hence, it is to be shown that

- K is reflexive and transitive
- E satisfies the conditions
  - *E1*: $\mathcal{C}(c) \subseteq E(w, \mathfrak{J}(c), \mathfrak{J})$
  - *E2*: $E(w,d_1,\mathfrak{J}) \circ E(w,d_2,\mathfrak{J}) \subseteq E(w, d_1 \cdot^D d_2, \mathfrak{J})$
  - *E3*: $E(w,d_1,\mathfrak{J}) \cup E(w,d_2,\mathfrak{J}) \subseteq E(w, d_1 +^D d_2, \mathfrak{J})$
  - *E4*: $\varphi \in E(w,d,\mathfrak{J})$ and $d = \mathfrak{J}(t) \Rightarrow t : \varphi \in E(w, !^D d, \mathfrak{J})$
  - *E5*: $\forall d(\varphi \in E(w, \mathfrak{J}_x^d[t], \mathfrak{J}_x^d)) \Rightarrow \forall x\varphi \in E(w, \mathfrak{J}[gen_x(t)], \mathfrak{J})$
  - *E6*: $\varphi \in E(w, \mathfrak{J}[t], \mathfrak{J}) \Rightarrow \forall v(R(w,v) \Rightarrow \mathfrak{M}, v \vDash \varphi)$
  - *E7*: $\mathfrak{J}_1 \overset{\varphi}{=} \mathfrak{J}_2 \Rightarrow \varphi \in E(w,d,\mathfrak{J}_1) \Leftrightarrow \varphi \in E(w,d,\mathfrak{J}_2)$
  - *EM*: $K(w,v) \Rightarrow E(w,d,\mathfrak{J}) \subseteq E(v,d,\mathfrak{J})$

Conditions *E1-E4* follows from **RJ** and axioms **J1**, **J2**, **J3**.

Condition *E5* follows from **JQ**.

Proof. Suppose $\varphi \in E(w, \mathfrak{J}_x^d[t], \mathfrak{J}_x^d)$, for all d. Then

$\varphi \in E(w, t_x^d, \mathfrak{J}_x^d)$, for all d, by lemma 1

$\varphi^d_x \in w/t^d_x$, for all d, by the deifintion of E
$(t:\varphi)^d_x \in w$, for all d, by the deifnition of substitution
$\forall x(t:\varphi) \in w$, since w is witnessed
$gen_x(t):\forall x\varphi \in w$, by axiom **JQ**
$\forall x\varphi \in w/gen_x(t)$, by the definition of $w/gen_x(t)$
$\forall x\varphi \in E(w, \mathfrak{I}[gen_x(t)], \mathfrak{I})$, by the definition of E □

Condition *E6* follows from **J**□.

Proof. Condition *E6*.
Suppose $\varphi \in E(w, \mathfrak{I}[t], \mathfrak{I})$. Then
$\varphi \in E(w, t, \mathfrak{I})$, by lemma 1
$t:\varphi \in w$, by the deifnition of E
$\Box\varphi \in w$, by axiom **J**□
$\varphi \in w/\Box$, by the definition of $w/\Box$
$\forall v(R(w,v) \Rightarrow \varphi \in v)$, by the definition of R
$\forall v(R(w,v) \Rightarrow \mathcal{M}, v \vDash \varphi)$, by the inductive hypothesis □

Lemma 3. *(Truth Lemma). Let* $M = \langle W, R, K, \mathfrak{D}, \mathfrak{I}, E, V\rangle$ *be the canonical model for* $\mathbf{QLJ_{S5}}$. *Then* $M, w \vDash \varphi \Leftrightarrow \varphi \in w$.

Proof. By cases.

(i) $t:\varphi \in w$.
$\varphi \in w/t$, by the definition of w/t
$\varphi \in E(w,t,\mathfrak{I}) = E(w, \mathfrak{I}[t], \mathfrak{I})$, by lemma 1
$\varphi \in w/K$ and $\varphi \in E(w, \mathfrak{I}[t], \mathfrak{I})$, by the definition of w/K
$\forall v(K(w,v) \Rightarrow \varphi \in v)$ and $\varphi \in E(w, \mathfrak{I}[t], \mathfrak{I})$, by the definition of K
$\forall v(K(w,v) \Rightarrow \mathcal{M}, v \vDash \varphi)$ and $\varphi \in E(w, \mathfrak{I}[t], \mathfrak{I})$, and so $M, w \vDash t:\varphi$

(ii): $t:\varphi \notin w$. Then $\varphi \notin E(w, \mathfrak{I}[t], \mathfrak{I})$, and so $M, w \nvDash t:\varphi$ □

4 Acknowledged existence of proofs

Let us extend our language by adding a new predicate $\mathbb{E}$ of existence, more precisely acknowledge existence. The system $\mathbf{QLJ_{S5E}}$ of quantified modal justification logic with existence is based on language $\mathcal{L}(\mathbb{E})$ is the extension of $\mathbf{QLJ_{S5}}$ obtained by adding the following axioms.

Group V: existence axioms

E1: $\mathbb{E}(t_1 \cdot t_2) \rightarrow \mathbb{E}(t_1) \wedge \mathbb{E}(t_2)$

E2: $\mathbb{E}(t_1 + t_2) \rightarrow \mathbb{E}(t_1) \wedge \mathbb{E}(t_2)$

E3: $\mathbb{E}(!t) \rightarrow \mathbb{E}(t)$

Intuitively, a proposition like $\mathbb{E}(t)$ states that the agent has acknowledged the existence of t, or that t is explicitly available to the agent. It is assumed that it is necessary for the agent to possess a proof t to construct every piece of evidence that constitutes t. Hence, axioms **E1-E3** are immediately justified in view of this assumption. The semantic framework is changed accordingly. In particular, a frame is a tuple $\langle W, R, K, \mathfrak{D}, \mathfrak{I}, \delta, E\rangle$ where

- $\langle W, R, K, \mathfrak{D}, \mathfrak{I}, E\rangle$ is a modal frame
- $\delta : W \rightarrow \wp(D)$ is a selection function assigning to each world $w \in W$ the set of justifiers available at w, and it is required to satisfy the following closure conditions:
  $\delta 1$: $d_1 \cdot^D d_2 \in \delta(w) \Rightarrow d_1, d_2 \in \delta(w)$
  $\delta 2$: $d_1 +^D d_2 \in \delta(w) \Rightarrow d_1, d_2 \in \delta(w)$
  $\delta 3$: $!^D d \in \delta(w) \Rightarrow d \in \delta(w)$

Finally, the truth definition of a formula like $\mathbb{E}(t)$ is given as follows

- $\mathcal{M}, w \vDash \mathbb{E}(t) \Leftrightarrow \mathfrak{I}(t) \in \delta(w)$

It is not difficult to see that conditions δ1-δ4 ensure the validity of axioms **E1-E4** and that the system is complete when δ in the canonical model is defined so that $\delta(w) = \{t \mid \mathbb{E}(t) \in w\}$.

4.1 Kinds of justification

Given the possibility of expressing actual proof possession, a number of different epistemic notions comes to be definable. In particular, with respect to justifiers, the following distinctions can be introduced.

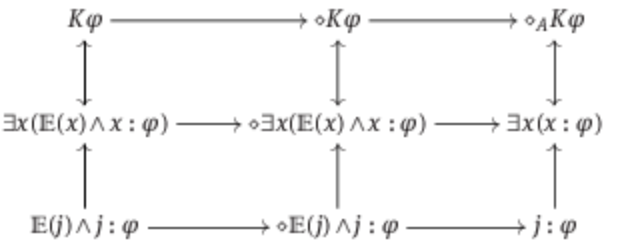
$$\mathbb{E}(j) \wedge j : \varphi \longrightarrow \Diamond \mathbb{E}(j) \wedge j : \varphi \longrightarrow j : \varphi$$

Hence, we are in a position to distinguish between

1. the fact that j justifies φ
2. the fact that it is possible to possess a j justifing φ
3. the fact that the agent actually possess a j justifing φ

4.2 Kinds of knowledge

Given the possibility of expressing actual proof possession, the notion K of actual knowledge turns out to be definable and, with respect to knoweldge, the following distinctions can be introduced, resting on the idea that knowledge is given by the existence of a justifier.



In this diagram, $K\varphi$ is a notion of actual knowledge, defined as possession of a justifier for φ; $\circ K\varphi$ is the corresponding notion of potential knowledge, defined as the possibility of possessing a justifier for φ ; and $\circ_A K\varphi$ is a notion of abstract possible knowledge, defined as the mere existence of a justifier for φ.

The foregoing distinctions are of interest in different context. In particular, they can be exploited for developing different kinds of analyses in epistemic logic and formal epistemology.[5]

5 Among the main areas of application, see Baltag et al. (2014); Dean and Kurokawa (2010); Duc (1997); Van Benthem and Velazquez-Quesada (2010).

Franz von Kutschera

The Case for Conceptualism

Abstract: The main tenets of this paper are: (1) The realistic conception of abstract objects is responsible for the logical and set theoretical paradoxes. (2) The usual objections against conceptualism are invalid. (3) Conceptualism is a suitable frame for theories of abstract objects.

1 Abstract objects

Abstract objects are concepts, propositions, sets, numbers, theories, ideals, poems, symphonies etc. They are distinguished from physical and psychological phenomena by not existing in time and therefore cannot be causes or effects. The topic of this paper is the ontological status of abstract objects. In effect it is the traditional problem of universals, even if I shall be more concerned with sets than with concepts.

The existence of abstract objects is denied by *nominalism*. Medieval nominalism was directed against essentialism, against real essences of things. It recognized only concrete individuals but nothing universal, hence its motto: *universale est vox*. A predicate like "red" does not express a property but is just a common, an ambiguous name – we cannot say "four red things", since this would presuppose the property of being red. For nominalists red things have nothing in common but being called "red" – well, that is another property, but you cannot think without conceiving and that means employing concepts.

At the beginning of modern philosophy Thomas Hobbes revived nominalism, though not in a strictly consequent way, since he assumed a similarity between the objects for which a common name stands.[1] Contemporary nominalism was developed mainly by Stanislaw Lesniewski, Nelson Goodman and David Lewis. It has little to do with the medieval idea. The nominalistic "Ersatz" for sets are wholes, consisting of parts. Their logic is mereology. Goodman called his mereological calculus the *calculus of individuals*, but for him individuals are not just concrete objects but anything that is a model of his calculus. Models are, for instance, complete Boolean algebras without a null element, which are

1 See *Leviathan*, ch. 4 and *De corpore* I, 2.9.

Franz von Kutschera:

DOI: 10.1515/9783110529494-014

very powerful systems of set theory. David Lewis has added a neighbourhood relation to the calculus of individuals as a basis for the definition of topological concepts. Modern nominalism, then, does not deny the existence of abstract objects anymore, it just uses another name for them.

2 The controversy about universals

The controversy about abstract objects is about their ontological status. There are mainly two positions:

- For *Platonism* or *Realism*, whose motto is: *universalia ante res*, they belong to a third realm of being besides the physical and the mental.
- For *Conceptualism*, whose motto is: *universalia in mente*, they are mental constructs.

The Aristotelian position that universals exist only in concrete things – *universalia in rebus* – has never been made sufficiently precise. If properties exist in things they are particulars, not universals. The idea was: Concrete things must have some trait that justifies the application of a universal like 'red'. This trait, however, would have to be at once particular and universal. Therefore I shall restrict my discussion of universals to realism and conceptualism.

3 Platonism

Because of the usual objections against conceptualism, which I shall discuss later, most logicians and mathematicians so far have been realists. Georg Cantor, one of the pioneers of set theory, was a realist just as Gottlob Frege[2] and Kurt Gödel. Gödel writes: "Classes and concepts may [...] be conceived as real objects [...] existing independently of our definitions and constructions. [...] It seems to me that the assumption of such objects is quite as legitimate as the assumption of physical bodies and there is quite as much reason to believe in their existence. They are in the same sense necessary to obtain a satisfactory theory of mathematics as physical bodies are necessary to obtain a satisfactory theory of sense perceptions".[3]

2 See Kutschera (1989), 10.2.

3 Gödel (1944), p. 137.

For Gödel mathematics is something like a natural science of the world of abstract objects, not a science of intellectual constructs. That was also Frege's position. For him abstract objects, concepts or propositions (*Gedanken*, as he said) and sets, were not constructed but discovered. In his paper *Der Gedanke* from 1918 he tries to show first that propositions are neither physical things nor mental ideas and then says: "The result, then, seems to be that propositions are neither things of the external world nor ideas (*Vorstellungen*). A third realm has to be recognized. What belongs to it equals ideas in that it cannot be perceived by the senses and concrete things in that it needs no subject to whose conscious life it belongs. The proposition, e.g., expressed by the theorem of Pythagoras, is true timelessly and independently of whether someone believes it to be true. It needs no subject. It is not only true since it has been discovered – rather it is like a planet that, long before it was discovered, has been acting on other planets and been acted upon by them."[4] And again: "In thinking we do not generate propositions but grasp them."[5] And: "Propositions are not psychological entities and thinking is not an inner production and construction but a grasping of propositions which are already there."[6]

As these quotations show Frege remained a Platonist as long as he lived. He never suspected that this position might be the origin of his troubles with set theory or even in any way problematic, although it is really much more natural to think of concepts and propositions as something we form instead of something we find out there in an intangible and invisible objective realm. As sets are extensions of concepts it would have been more natural to think of them also as generated together with the concepts.

4 Epistemological problems

Platonism, first of all, has serious epistemological problems. It proposes a realistic conception of abstract objects and the states of affairs concerning them – let me call them *abstract states* – and that would have to be defined by their analytical independence of mental states of affairs, of our thoughts, perceptions and imaginations. The intuition is that you cannot infer mental states from mathematical ones, and vice versa; each possible abstract state is compatible with each possible mental state.

4 Frege (1967), p. 353 f.
5 Frege (1967), p. 359.
6 Frege (1976), p. 102.

This description, however, becomes trivial if abstract states are analytic, for then all possible abstract states hold in all possible worlds and therefore are compatible with all other possible states. Thus a realist has to assume that abstract states are synthetic. Logical and mathematical truths, however, can be recognized a priori, i.e. independently of empirical observations. Aside from very few authors as John Stuart Mill nobody considers mathematical truths as empirical and maintains that the statement "5+4 = 9" is an inductive generalization of observations that 5 apples and 4 apples are 9 apples, 5 cherries and 4 cherries are 9 cherries, etc., and could be falsified by future observations. Only very few authors like Kant and Fichte admit synthetic truths a priori and for them they are about the mental, about our ways of thinking and perceiving. The result, then, is this: Since abstract truths can be know a priori they are about the mental or they are analytic. In both cases they cannot be conceived realistically.

At least in simple logical and mathematical statements we are free from error. But as Frege said: "With the step into an external world I expose myself to error."[7] The world of abstract objects cannot then be an external world as Platonism has it.

5 Logical paradoxes

I do not want to go into these epistemological problems of Platonism any deeper, the main theses of my paper is rather: *The realistic conception of abstract objects is responsible for the logical and set theoretical paradoxes*.

From the beginning of his work in logic Frege's central intention was to clear up the meaning and the basis of number theoretic propositions.[8] With his set-theoretical foundation of arithmetic in the first volume of the *Grundgesetze* which appeared in 1893 – the main ideas are already contained in the *Grundlagen der Arithmetik* from 1884 – he seemed to have reached his goal. This book was the outcome of unusually penetrating, careful and far reaching considerations, formulated in a formal system of unequalled exactness. In it Frege had proven a great number of theorems of logic, set theory and arithmetic. He even tried to show (in §31), that each well-formed expression of his formal language has exactly one extension. If this proof were correct it would establish the consistency of the system.

All this was ample legitimation for Frege's claim in the introduction to the first volume of his *Grundgesetze*: "It is quite unlikely that such a construction

7 See Frege (1967), p. 358.

8 Frege (1976), pp. 359 f and Frege (1983), p. 282.

could be erected on a defective and uncertain ground. [...] And only a [similarly] detailed demonstration that on other fundamental assumptions a better and firmer building could be constructed I should accept as a refutation of my theory, or if someone should show that my principles lead to evidently wrong consequences. But that nobody will achieve." (p. XXVI)

Already in the appendix of the second volume of the same work from 1903, however, we read: "For a scientific author there is nothing more disagreeable than if, after the conclusion of his work, one of its foundations is upset. Into this situation I was put by a letter of Mr. Bertrand Russell when the present work was already being printed." (p. 253) In an appendix Frege could only briefly point out the contradiction, Russell's paradox, and make a proposal how to avoid it. This proposal has later been shown to be ineffective by St. Lesniewski in 1938 an by W.V. Quine in 1955, but from the beginning it was no more than a hasty attempt to patch the leak, an attempt which Frege himself did not take seriously afterwards; he never referred to it again.

Georg Cantor, a contemporary of Frege, never set up a theory of sets satisfying modern criteria of exactness as they were laid down by Frege. He, therefore, could view the set theoretical paradoxes he had discovered already before 1902 as strange but only ephemeral phenomena. For Frege, however, they were plainly and simply a catastrophe. For according to the principle *ex falso quodlibet* a single contradiction implies everything, so that a contradiction is not just a local breakdown but a global catastrophe. It is characteristic for Frege's intellectual honesty that he did not try to deceive either himself or others about this disaster. For the rest of his life he tried in vain to overcome it. In his letter to Russell of June 22th he writes: "Anyway, your discovery is very strange and will perhaps initiate a great progress in logic, annoying as it seems at first glance."[9] As clear-sighted as this conjecture proved to be, Frege himself had no part in this progress. He never found "the right point of view", as he wrote in the same letter, and so the rapid development of logic and set theory after the turn of the century, for which he had laid the foundations, went on without him.

In his *Grundgesetze der Arithmetik* Frege does not speak of sets but of courses of values of functions. As this is just a minor point I will ignore it here. Frege's concept of a set is the classical one of an extension of a predicate. The set of frogs is the extension of the predicate "frog", and it has as elements exactly those things that are frogs. Frege's set theory – or "naive set theory" as hindsight has it – is defined by two simple axioms that are direct consequences of this concept. The axiom of *comprehension* says: *Every (one-place, 1st order) predicate has an*

9 Frege (1976), p. 215.

extension. The axiom of *extensionality* is: *Sets that have exactly the same elements are identical*. These two principles seem to be quite trivial and harmless as they just explicate the notion of a set.

Nevertheless they give rise to Russell's paradox and a number of other contradictions. Russell's set is the set of all sets that are not elements of themselves. It is an element of itself if and only if it is not an element of itself. In fact the paradox follows from the comprehension axiom alone. But where is the mistake in that simple principle? For someone as Frege who had recognized set theory as the basis of mathematics this was a tormenting and deeply frustrating problem. He was acquainted with the attempts of Russell and Ernst Zermelo to weaken set theory so that the known contradictions could not be derived anymore in the usual way. In type-theoretical systems as well as in those of axiomatic set theory the classical comprehension principle is restricted. Frege rightly thought all of them ad hoc and intuitively unconvincing. Type-theory, he saw, makes sense only as a higher predicate-logic and as such Frege himself had employed it in his *Begriffsschrift* from 1879.[10] Systems of axiomatic set theory as the standard *ZFF* from the beginning of the 1920s, on the other hand, try to restrict set formation only so far that it still yields the sets necessary for the mathematically important theories of cardinal and ordinal numbers. Such systems, then, are chosen from purely pragmatic aspects and remain intuitively unconvincing. They are a far from the ideal to derive the basic principles of set theory from an intuitively well-defined concept of set and do not identify the origin of the paradoxes. That, I maintain, is the realism that seems to come so naturally to mathematicians and that is still the basis of the (simple) theory of types and axiomatic set theory. Realism has to take the universe of all existing sets as given and therefore has no reason whatsoever not also to assume the existence of Russell's set. Thus it has no acceptable way out of his paradox.

6 Objections to conceptualism

While Platonism cannot account for the a priori evidence of logical and mathematical statements conceptualism can do so since the nature of our own constructs cannot be hidden from us. A conceptualistic notion of concepts and

10 Frege's letter to Russell from September 23^{th} 1902, Frege (1976), p. 277 f. Against Russell's suggestions (see. his letter of 8.8.1902 in Frege (1976), p.226) Frege insisted that sets are objects, and that there is no natural basis for a distinction of sorts or levels of objects as there is in the realm of concepts. (See. Frege's answer to Russell's letter of 23.9.1902 in Frege (1976), p.227 f.)

propositions was first developed by the Stoics. According to them concepts and propositions belong to the realm of the mental. They are constructs of our thinking, formed by us and not discovered, as Frege believed. The usual objections to this notion are:

1) Mental states and acts are states and acts of a certain subject while concepts have no subject.

2) A concept as that of prime number is timeless. If it were a mental construct we would have to say that it probably came into being in the 6^{th} century B.C. in Babylonian mathematics.

The answer to both objections is that we have to distinguish between tokens and types, between particular acts and act types, between John's conceiving of a certain object as a frog on a certain day and concepts as forms of such acts. Types of acts have no individual subjects and no situation in time.

7 A constructivistic conceptualism

For a satisfactory conceptualistic theory of abstract objects we have to construct these objects in a systematic way. A concept, first of all, is a type of conceiving objects in a certain way, e.g. as green or as a frog. We can also reflect on concepts as ways of conceiving and make them the object – in the wider sense of a topic – of considerations and characterize them by using other concepts. Reflection does not make objects in the narrow, categorial sense out of con-cepts, of course. They remain something that can also be used and applied. On a first level concepts are applied to individuals of a set U. If we reflect on them and consider them as new objects our universe of discourse is enlarged and we may form new concepts, concepts of a second level such as 'is a transitive relation' or 'is a property of John'. We may then iterate this step and pass on to a still larger domain containing also the concepts of level 2 as objects and define concepts of level 3 on it. In this way we can construct an open hierarchy of concepts and their domains, but we never arrive at a set of all concepts.

This constructive approach to concepts arises from the basic idea of conceptualism, from the ideas of concepts as types of mental acts, and does not fit into a realistic notion of concepts. For a Platonist there is the eternal set of all objectively existing concepts and an all including universe of discourse. Therefore there is also the property of being a property that does not apply to itself. Clearly this property applies to itself if and only if it does not apply to itself. This paradox

cannot be constructed in the conceptualistic hierarchy of concepts since it yields no concepts that are defined for themselves.

For conceptualists there is a corresponding hierarchy of propositions: Propositions of level 1 are those that are not about propositions, propositions of higher levels can also be about those of a lower level. In such a hierarchy Tarski's paradox of self-applying propositions cannot be constructed:[11] A universal proposition of the form 'All propositions have the property E' – for instance: 'All propositions maintained by Max are false' – is called "self-applying" if it has property E itself, in our example: if the proposition is not maintained by Max or is wrong. Is the proposition, that all propositions are not self-applying, self-applying or not? It is self-applying if and only if it is not self-applying. This paradox has the same structure as that of the self-applying concepts and Russell's paradox. We don't need an axiom of comprehension for it but just the realistic assumption that all objectively existing universal propositions form a single given class. Then in this class there is the difference of self-applying and not self-applying propositions, and the proposition that all of them are not self-applying is already among them. A conceptualist, however, will say: If we have a set $P1$ of propositions we may make statements about all propositions in $P1$ and form appropriate propositions. But since forming them, we presuppose the elements of $P1$ they cannot be among these elements. If we enlarge the set $P1$ to $P2$ by adopting the new elements the word "all" changes its meaning if applied to the elements of the larger set $P2$ instead of to $P1$. The proposition that all propositions of $P1$ are not self-applying does not belong to $P1$. It is not about itself, and therefore the contradiction vanishes.

8 A conceptualist set theory (1st version)

How a conceptualistic theory of sets may look is shown by the theory George Boolos developed in his paper *The iterative concept of set* (1971). His notion of a set is conceptualistic: Sets are not given but formed. They are formed by collecting already existing objects. Cantor wrote: "By a 'set' we understand every collection M of well-defined and clearly distinguished objects of perception or thought (called 'elements' of M) into a whole."[12]

I shall describe the theory only intuitively. The process of set-formation starts with a class, V_0, of given individuals. In pure set theory this class will be empty,

11 Tarski has formulated it for sentences, see Tarski (1949), p. 80, footnote 11.

12 Cantor (1895-1897), §1.

but for the heuristic exposition let us first suppose that there are individuals. In a first step we can form arbitrary sets of objects from V_0. These sets are new objects so that we can enlarge V_0 to a class V_1 that contains all collections of elements of V_0. If $\wp(X)$ is the power set of X, we then have

a) $V_1 = V_0 \cup \wp(V_0)$.

This step may be repeated, so that we generally have

b) $V_{n+1} = V_n \cup \wp(V_n)$.

If the V_n are defined in this way for all natural numbers n, i.e. for all finite ordinals, we can, in a next step, form the set V_ω as the union of all the V_n's – ω being the smallest transfinite ordinal. Therefore we set

c) $V_\omega = \bigcup_{\alpha<\omega} V_\alpha$.

Thus in V_ω as the universe of discourse we obtain no new objects. New objects come up only if we move on from V_ω to $V_{\omega+1}$, and so on. Let $\alpha, \beta, \gamma, \ldots$ be ordinals, and $\lambda, \lambda', \ldots$ limit numbers, i.e. ordinals that have no immediate predecessor. Let us further assume now that V_0 is empty. Then we have

d) $V_{\alpha+1} = \wp(V_\alpha)$ and $V_\lambda = \bigcup_{\alpha<\lambda} V_\alpha$.

The result is the well-known cumulative von-Neumann hierarchy of sets. If $O(x)$, the *order* of set x, is the smallest ordinal α such that $x \in V_\alpha$, we have $x \in y \rightarrow O(x) < O(y)$. Therefore all sets are *grounded*, i.e. there is no infinite sequence of (not necessarily different) sets in which each one contains the next as an element. Each iterative notion of set in the general sense that sets are constructed step by step, where every step presupposes the existence of the objects already constructed, gives rise only to grounded sets.

Since ordinal numbers are to be introduced only in set theory, we cannot already use them in the axioms of such a theory. Therefore these considerations have only a heuristic function. I cannot define an axiomatic theory here, but Boolos has shown that Zermelo-Fraenkel set theory (with the axiom of foundation, i.e. *ZFF*) is an appropriate system for his iterative concept of set. This can best be seen by substituting Dana Scott's equivalent system Σ for *ZFF*.[13]

13 See Scott (1974).

Boolos' conception of sets as collections is not very convincing, however. We know how to collect physical things like stamps or coins, although not into a "whole" but only into some sort of container, but how do we collect abstract objects? And how can we collect infinitely many objects? How can a new object – the empty set – be generated by collecting nothing? Or something different from a penny – namely the unit set of this penny – by collecting just one penny? And why should the collection of two collections be different from what you get by simply collecting their elements?

9 A conceptualist set theory (2nd version)

Intuitively the classical notion of set as an extension of a predicate is much more convincing than that of a collection.[14] According to it sets are introduced relatively to a set U of given objects by abstraction in the form $\{x \in U : A[x]\}$ – the set of all $x \in U$ to which the predicate $A[x]$ applies. The quantifiers in this predicate have to be restricted to U also, so that no other objects are assumed, and the names must stand for objects from U. Again I sketch this way of building a hierarchy of sets only intuitively: We begin with a set U_0 of individuals. If all quantifiers in $A[x]$ are restricted to U_0, we create a new object $\{x \in U_0 : A[x]\}$ as the extension of the predicate $A[x]$ on U_0. Let the set of sets we can construe in this way be $D(U_0)$ then U_1 is to be $D(U_0) \cup U_0$. The next sets $U_{\alpha+1}$ are defined in the same way. If λ is a limit ordinal we have $U_\lambda = \bigcup_{\alpha<\lambda} U_\alpha$. The difference to our first approach is that we use $D(U_\alpha)$ instead of the power set $\wp(U_\alpha)$. Up to U_ω there is no difference, but generally $D(U_\alpha)$ is a proper subset of $\wp(U_\alpha)$. $D(U_\alpha)$ may be defined by set operations as Gödel has shown in (1940).[15]

If U is the union of all U_α, it is the class of constructible sets. Again we can assign every $x \in U$ an order $Or(x)$, the smallest ordinal α with $x \in U_\alpha$. For $x \in y$ we then have $Or(x) < Or(y)$. It is not generally true anymore, however, that $Or(y) \leq \alpha$, if for all x such that $x \in y$ we have $Or(x) < \alpha$, as for the the collective concept of set. There $O(x)$ is the smallest ordinal greater than every order of an element of x. Here we can only say that $Or(x)$ is greater than all the orders of elements of x. There may be subsets of U_α which are not elements of $U_{\alpha+1}$ but only elements of some U_γ for which γ is much larger than $\alpha + 1$.

All sets in U are grounded again. If $r(x, y)$ is the relation on U holding if the construction of y presupposes the existence of x this relation is grounded. In our

14 See Kutschera (2001).

15 See Jech (2003), pp. 175 ff.

second version of set theory, however, $r(x, y)$ does not hold only if x is connected by an $\in$-chain with y, i.e. if there are $z_1, \ldots, z_n$ with $x \in z_1, z_1 \in z_2, \ldots, z_n \in y$. The construction of y may also presuppose sets that are neither elements of y nor elements of elements of y, and so on.

We also need principles for the existence of U_α-sets. The most attractive one is the power-set axiom:

$$PA\colon x \in U \to \wp(x) \in U.$$

It is not provable in the theory of constructible sets but consistent with it. If we add it this theory is equivalent to *ZFF*. All the necessary considerations are due to Gödel.

In this paper I have not been concerned with formal systems but only with the idea that the only satisfactory approach to a theory of abstract objects is a constructive conceptualism and that it is also the only satisfactory way of eliminating the set-theoretical paradoxes. Sets are not something discovered in a Platonic sky but mental constructs. We have to build them step by step from individuals, sets of individuals, and so on. This yields a hierarchy of sets in which the paradoxes cannot be construed anymore. This hierarchy is open in the sense that it may be continued indefinitely. There is neither a set of all sets nor a set of all concepts or a set of all propositions.

Wolfgang Lenzen

Two days in the life of a genius

1 Introduction

Gottfried Wilhelm Leibniz (1646–1716) is generally regarded as the last "universal scientist" who exhibited an extraordinary knowledge in almost all fields of science. In particular, he was a brilliant mathematician who, independently of Newton, invented the infinitesimal calculus and who discovered the importance of the dyadic number system as a basis for calculating machines. When he died at the age of 70, he left behind an extensive and widespread collection of papers on topics ranging from Theology, Jurisprudence, Medicine, Philosophy, Philology, Geography and all kinds of historical investigations to Mathematics, the Natural Sciences and some less scientific matters such as the Military and the Foundation of Societies and Libraries. But throughout his life, Leibniz didn't publish a single paper on *logic*, except perhaps for the mathematical dissertation "De Arte Combinatoria" or the juridical disputation "De Conditionibus". The former incidentally deals with some issues in the traditional theory of the syllogism, while the latter contains some interesting observations about the validity of principles of what is nowadays called deontic logic. Leibniz's main aim in logic, however, was to extend Aristotelian syllogistic to a "Calculus Universalis". Although there exist several drafts for such a calculus which seem to have been composed for publication, Leibniz appears to have remained unsatisfied with these attempts. Anyway he refrained from sending them to press.

In June 1690, after having spent almost three years in Bavaria, Austria, and Italy where he completed the task of inquiring the origin of the House of Welf, Leibniz returned to Hanover. Not much is known about his scientific activities during that time. In the chronicle of Leibniz's life and work, it is only reported that in July he solved a mathematical problem put forward by J. Bernoulli, and that in the first two days of August he composed two papers entitled "Primaria calculi logici fundamenta" and "Fundamenta calculi logici"[1].

1 Cf. Müller and Krönert (1969, p. 104–105).

Wolfgang Lenzen: ...

DOI: 10.1515/9783110529494-015

The aim of the present contribution is to scrutinize these short but very sophisticated papers in order to show that Leibniz was not only a mathematical mastermind, but also a genius logician.

2 Systematic background

In order to facilitate the understanding of Leibniz's logical investigations, let us briefly summarize the main elements of his logic as they have been developed above all in the "General Inquiries" of 1686 (*GI*, for short)[2].

2.1 The Algebra of concepts, L1

This calculus presupposes a potentially infinite set of monadic predicates or concepts $A, B, C \ldots$ for which the following operators are defined:

$A = B$ "A is the same as B"; "A coincides with B"

$A \in B$ "A contains B"; "A is B"

AB the conjunction of A and B

$\sim\sim A$ the negation of A

$P(A)$ "A is possible"; "A is 'a thing'"; "A is 'being'"

Identity or "coincidence" of concepts may be axiomatized by the law of *reflexivity*,

Iden 1: $A = A$ (*GI*, §171),

in conjunction with the rule of *substitutivity*: "That A is the same as B means that one can be substituted for the other in any proposition without loss of truth"[3], i.e.

Iden 2: If $A = B$, then $\alpha[A] \leftrightarrow \alpha[B]$.

By means of these two principles, one easily derives the following corollaries:

Iden 3: $A = B \rightarrow B = A$
Iden 4: $A = B \wedge B = C \rightarrow A = C$

2 Cf. the text-critical edition in *Acad* VI, 4, 739–788; an English translation may be found in *LLP*, 47–87.

3 Cf. *LLP*, 52; cf. also *Cout*, 362: "Idem autem esse A ipsi B significat alterum alteri substitui posse in propositione quacunque salva veritate".

Iden 5: $A = B \rightarrow \sim A = \sim B$
Iden 6: $A = B \rightarrow AC = BC$.[4]

Let it be noted in passing that identity might as well be *defined* by means of the operator of conceptual containment, $\in$, according to the law:

Iden 7: $A = B \leftrightarrow A \in B \wedge B \in A$.[5]

Conversely, the relation $\in$ can be defined in terms of = according to the law "Generally 'A is B' is the same as 'A = AB'" (*GI*, §83):

Cont 1: $A \in B \leftrightarrow A = AB$.

The following two laws express the *reflexivity* and the *transitivity* of the $\in$-relation: "B is B" (*GI*, §37); "If A is B and B is C, A will be C" (*GI*, §19), i.e.:

Cont 2: $A \in A$
Cont 3: $A \in B \wedge B \in C \rightarrow A \in C$.

The most fundamental principle for the operator of conceptual *conjunction* says: "That A contains B and A contains C is the same as that A contains BC" (*GI*, §35), i.e.:

Conj 1: $A \in BC \leftrightarrow A \in B \wedge A \in C$.

Conjunction or "addition", as Leibniz sometimes also calls it, is *idempotent* and *symmetric*: "AA = A" (*GI*, §171); "AB = BA" (*Cout*: 235, ¶7):

Conj 2: $A = AA$
Conj 3: $AB = BA$.

Some further theorems comprise:

Conj 4: $AB \in A$ "AB is A" (*Cout*: 263).
Conj 5: $AB \in B$ "AB is B" (*GI*, §38).

4 Cf. *GI* §6: "If A coincides with B, B coincides with A"; §8: "If A coincides with B and B coincides with C, then A also coincides with C"; §2: "If A and B coincide, so also do Not-A and Not-B"; §171: "Fifth, if A = B, AC = BC".

5 Cf. *GI* §30: "That A is B and B is A is the same as that A and B coincide".

The next operator is conceptual *negation*, 'not'. Leibniz had serious problems with finding the proper laws governing this operator. From the tradition, he knew little more than the "law of double negation":

Neg 1: $\sim\sim A = A$ "Not-Not-A = A" (*GI*, §96).

One important step towards a complete theory of conceptual negation was to transform the informal principle of contraposition, 'Every A is B, therefore every Not-B is Not-A' into the following abstract axiom:

Neg 2: $A \in B \rightarrow \sim B \in \sim A$.[6]

Furthermore Leibniz discovered various variants of the "law of consistency":

Neg 3: $A \neq \sim A$
Neg 4: $A = B \rightarrow A \neq \sim B$.

In the *GI* these principles (which are easily seen to be equivalent to each other) are formulated as follows: "A proposition false in itself is 'A coincides with Not-A"' (§11) and Seventh: "If A = B, then A ≠ Not-B" (§171).

Similarly, the subsequent two principles are provably equivalent to each other:

Neg 5*: $A \notin \sim A$
Neg 6*: $A \in B \rightarrow A \notin \sim B$.[7]

These principles have been marked with a '*' to indicate that they are not absolutely valid but have to be restricted to self-consistent terms!

This remarks leads us to the last operator of L1, conceptual self-consistency or *possibility*, $P(A)$, which can be defined by the following equivalence:

Poss 1: $P(A) \leftrightarrow A \notin (B \sim B)$.

Thus in *GI* §§32–33 Leibniz explains: "B Not-B is impossible; or, if B Not-B = C, C will be impossible. So if A = Not-B, AB will be impossible." Leibniz used to express the impossibility of a concept A also by calling it "not being" ("non Ens") or not

6 *GI* §77: "In general, 'A is B' is the same as 'Not-B is Not-A"'.

7 Cf. *GI* §43: "It is false that B contains Not-B, i.e. B doesn't contain Not-B"; and §91: "A is B, therefore A isn't Not-B".

being a "thing" ("non est res"). E.g., in §171 of *GI* he states: "Eighth, A Not-A is not a thing", i.e.

Poss 2: $\neg P(A \sim A)$.

A characteristic property of the operator P is expressed by the law:

Poss 3: $A \in B \wedge P(A) \rightarrow P(B)$.[8]

Furthermore, there exists a fundamental relation between the containment operator $\in$ and the impossibility of a corresponding complex concept:

Poss 4a: $A \in\sim B \leftrightarrow \neg P(AB)$
Poss 4a: $A \in B \leftrightarrow \neg P(A \sim B)$.

Thus in *GI*, §200 Leibniz notes: "If I say 'AB does not exist', this is the same as if I were to say 'A contains Not-B' [...]. Similarly, if I say 'A Not-B does not exist', this is the same as if I were to say [...] 'A contains B' ".[9] With the help of the operator 'P', the earlier principles **Neg 5*** and **Neg 6*** can be corrected as follows:

Neg 5: $P(A) \rightarrow A \notin\sim A$
Neg 6: $(P(A) \rightarrow A \in B \rightarrow A \notin\sim B)$.

A complete axiomatization of $L1$ may be obtained by adding the following counterpart of the "ex contradictorio quodlibet":

Neg 7: $(A - A) \in B$.

As a corollary of **Neg 7**, the impossible concept $A \sim A$ contains besides B also $\sim B$ and hence $(B \sim B)$, just as conversely $(B \sim B) \in (A \sim A)$, i.e. $(A \sim A) = (B \sim B)$. In other words, two self-inconsistent concepts necessarily coincide:

Poss 5: $\neg P(A) \wedge \neg P(B) \rightarrow A = B$.

---

8 Cf. *GI* §55: "If A contains B and A is true, B is also true". The variant $A \in B \wedge \neg P(B) \rightarrow \neg P(A)$ is formulated, e.g., in *Acad* VI, 4, 935: "Not-being or impossible is what involves a contradiction as A = BC Not-C. Hence, what involves something impossible is itself impossible".

9 Cf. also *Cout* 407/408: " 'A contains B' is a true proposition if A not-B entails a contradiction. [...] Hence it follows that A not-B implies a contradiction if 'A is B' is a true proposition, because for 'A' one can substitute the equivalent term 'AB' and obtain 'AB not-B' what is a manifest contradiction."

To conclude our discussion of $L1$, let it be pointed out that Leibniz frequently made the gross mistake of equating 'A isn't B' and 'A is Not-B'. Again and again he assumed that the following principle is correct:

Neg 8*: $A \notin B \leftrightarrow A \in\sim B$.

Thus in a preliminary version of §33 *GI*, he put forward the "Axiom: 'A does not contain B' is the same as 'A contains Not-B'".[10] And in §82 he maintained once more that "'B isn't A' is the same as 'B is Not-A', therefore 'B ≠ AY' is the same as 'B = Y Not-A'." As was argued above in connection with **Neg 6***, one "half" of the equivalence **Neg 8*** is "almost" correct, i.e. $A \in\sim B$ entails $A \notin B$ provided that A is *self-consistent*. However, this additional premise $P(A)$ doesn't suffice to warrant the validity of the other "half" of **Neg 8***! As will turn out in section 2.2 below, the inference $A \notin B \leftrightarrow A \in\sim B$ holds only in the rare case where A is an *individual concept*!

Thus in a paper of around 1685, he put forward a "*Second Rule of contradiction*: the truth of 'A is the same as B' coincides with the falsity of 'A is the same as Not-B'."[11] Also in the ripe paper "Fundamenta Calculi Logic" of 1690 which will be analyzed in more detail below, a first version of principle (14) said: "'A = B' is equivalent to 'A ≠ Not-B'." Of course, one "half" of the equivalence **Neg 9*** is correct, since according to **Neg 4**, $A =\sim B$ entails $A \neq B$; but the other "half" is so-to-speak "absolutely" wrong; the inference from $A \neq B$ to $A =\sim B$ would not even hold under the additional assumption that A is an individual concept.[12]

2.2 The Quantifier Logic, L2

Leibniz's quantifier logic $L2$ results from $L1$ by the introduction of so-called "indefinite concepts". These concepts are symbolized by letters from the end of the alphabet $X, Y, Z, \ldots$, and they function as *quantifiers* ranging over *concepts*. Thus in the *GI* Leibniz explains:

> (16) An affirmative proposition is 'A is B' or 'A contains B' [...]. That is, if we substitute the value for A, one obtains 'A coincides with BY'. For example, 'Man is an animal', i.e. 'Man'

10 Cf. *Acad* VI, 4, p. 753, text-critical apparatus to line 21.

11 Cf. *Acad* VI, 4, 624.

12 In Lenzen 1986 it is shown that in his "darker moments" Leibniz again and again resorted to **Neg 8*** and **Neg 9*** although in other moments he had clearly recognized the invalidity of these principles.

> is the same as 'a … animal' (namely, 'Man' is 'rational animal'). For by the sign 'Y' I mean something undetermined, so that 'BY' is the same as 'Some B', or 'a … animal' […], or 'A certain animal'. So 'A is B' is the same as 'A coincides with some B', i.e. 'A = BY'.

With the help of the modern symbol for the existential quantifier, the latter law can be expressed more precisely as follows:

Cont 4: $A \in B \leftrightarrow \exists Y(A = BY)$.

As Leibniz himself noted, the formalization of the UA according to **Cont 4** is provably equivalent to the simpler representation according to **Cont 3**:

> It is noteworthy that for 'A = BY' one can also say 'A = AB' so that there is no need to introduce a new letter.[13]

On the one hand, according to the rule of existential generalization,

Exist 1: If $\alpha[A]$, then $\exists Y \alpha[Y]$.

$A = AB$ immediately entails $\exists Y(A = YB)$. On the other hand, if there exists some Y such that $A = YB$, then according to **Iden 6**, $AB = YBB$, i.e. $AB = YB$ (by **Conj 2**) and hence (by the premise $A = YB$) $AB = A$.[14]

Next observe that Leibniz often used to formalize the particular affirmative proposition 'Some A is B' by means of the indefinite concept Y as '$YA \in B$'. In view of principle **Cont 4**, this representation can further be transformed into the (elliptic) equation $YA = ZB$. However, both formalizations are *inadequate* because they are easily shown to be *provable* in *L2*! According to **Cont 4**, $BA \in B$, hence by **Exist 1**:

Conj 6: $\exists Y(YA \in B)$.

Similarly, according to **Cont 3**, $AB = BA$, hence a twofold application of **Exist 1** yields:

Conj 7: $\exists Y \exists Z(YA = BZ)$.

---

13 Cf. *Cout* 366; cf. also *LLP* 56, fn. 1.

14 Exactly the same proof was given in "Primaria Calculi Logici Fundamenta" to be examined below.

These *tautologies*, of course, cannot adequately represent the particular affirmative proposition 'Some A are B'. Similarly, the corresponding formulas $\exists Y(YA \in \sim B)$ and $\exists Y \exists Z(YA = Z \sim B)$ fail to constitute correct formalizations of the particular negative proposition 'Some A are not B'.

In order to resolve these difficulties, let us consider a draft of a calculus probably written between 1686 and 1690[15], where Leibniz proved principle

Neg 10*: $A \notin B \leftrightarrow \exists Y(YA \in \sim B)$.

Leibniz's proof of this important law is really remarkable:

> (18) [...] 'A isn't B' is the same as to say 'there exists a Y such that YA is Not-B'. If 'A is B' is false, then 'A Not-B' is possible by [POSS 4A]. 'Not-B' shall be called 'Y'. Hence YA is possible. Hence YA is Not-B. Therefore we have shown that, if it is false that A is B, then QA is Not-B. Conversely, let us show that if QA is Not-B, 'A is B' is false. For if 'A is B' would be true, 'B' could be substituted for 'A' and we would obtain 'QB is Not-B' which is absurd.

To conclude the sketch of $L2$, let us consider some of the rare passages where an indefinite concept functions as a universal quantifier. In a draft of a calculus probably written between 1686 and 1690, Leibniz put forward the principle "'A is B' is the same as 'If L is A, it follows that L is B'":

Cont 5: $A \in B \leftrightarrow \forall Y(Y \in A \rightarrow Y \in B)$.

Furthermore, in §32 *GI*, Leibniz at least vaguely recognized that just as (according to CONJ 6) $A \in B$ is equivalent to $\exists Y(A = YB)$, so the negation $A \notin B$ is equivalent to the condition that, for any indefinite concept Y, $A \neq BY$:

Cont 6: $A \notin B \leftrightarrow \forall Y(A \neq YB)$.[16]

With the help of the universal quantifier '$\forall$', Leibniz's conception of individual concepts as maximally-consistent concepts can be formalized as follows:

Ind 1: $Ind(A) \leftrightarrow_{df} P(A) \wedge \forall Y(P(AY) \rightarrow A \in Y)$.

---

15 Cf. *Cout* 259–261, or the text-critical edition in *Acad* VI 4, 171. The editors of *Acad* guess that the paper was written around 1686

16 Cf. *Acad* VI, 4, 753: "(32) Propositio Negativa. A non continet B, seu A esse (continere) B falsum est, seu A non coincidit BY". The last passage "seu A non coincidit BY" had been overlooked by Couturat and it is therefore also missing in Parkinson's translation!

Thus A is an individual concept iff A is self-consistent and A contains every concept Y which is compatible with A.[17] Note, incidentally, that **Ind 1** might be simplified by requiring that, for each concept Y, A either contains Y or contains $\sim Y$:

Ind 2: $Ind(A) \leftrightarrow \forall Y(A \in\sim Y \leftrightarrow A \notin Y)$.

As a corollary it follows that the invalid principle **Neg* 8** holds only when it is restricted to individual concepts:

Neg 8: $Ind(A) \rightarrow (A \notin B \leftrightarrow A \in\sim B)$.

Already in the "Calculi Universalis Investigationes" of 1679, Leibniz had pointed out:

> [...] if two propositions are given with exactly the same singular [!] subject, where the predicate of the one is contradictory to the predicate of the other, then necessarily one proposition is true and the other is false. But I say: *exactly the same [singular] subject*, for example, 'This gold is a metal', 'This gold is a not-metal'.[18]

The crucial issue here is that **Neg* 8** holds only for individual concepts as, e.g., 'Apostle Peter', but not for *general concepts* as, e.g., 'man'. The text–critical apparatus of the Academy–edition reveals that Leibniz was somewhat diffident about this decisive point. He began to illustrate the above rule by the correct example "if I say 'Apostle Peter was a Roman bishop', and 'Apostle Peter was not a Roman bishop'" and then he went on, erroneously, to generalize this law for arbitrary terms: "or if I say 'Every man is learned' 'Every man is not learned"'. Finally he noticed this error "Here it becomes evident that I am mistaken, for this rule is not valid."

17 The underlying idea of the completeness of individual concepts has been formulated in § 72 *GI* as follows: "So if BY is ["being"], and the indefinite term Y is superfluous, i.e., in the way that 'a certain Alexander the Great' and 'Alexander the Great' are the same, then B is an individual. If the term BA is ["being"] and if B is an individual, then A will be superfluous; or if $BA = C$, then B=C." For a closer interpretation of this idea cf. Lenzen 2004.

18 Cf. *Acad* VI, 4, 217–218; a discussion of this important passage may be found in Lenzen (1986, pp. 23–24).

2.3 Formal Representations of the Categorical Forms

The traditional theory of the syllogism may be considered as the logic of the four categorical forms of a universal affirmative (UA), universal negative (UN), particular affirmative (PA), and particular negative (PN) proposition:

UA	**Every *A* is *B***	**UN**	**No *A* is *B***
PA	**Some *A* is *B***	**PN**	**Some *A* isn't *B***

These propositions can be represented in Leibniz's logic in various ways. In particular, one obtains the following "homogenous" formalization in terms of the operator $\in$ which will be referred to as *Schema 1*:

$A \in B$	$A \in \sim B$
$A \notin \sim B$	$A \notin B$

The homogeneity consists in two facts:

(i) The formula for the UN is obtained from the UA by just replacing predicate B with $\sim B$; this is the formal counterpart of the traditional principle of obversion according to which 'No A is B' is equivalent with 'Every A is not-B'.

(ii) In accordance with the traditional laws of opposition, the formulas for the particular propositions are just the negations of the »opposite« universal propositions.

In view of **Cont 1**, *Schema 1* may be transformed into *Schema 2*:

$A = AB$	$A = A \sim B$
$A \neq A \sim B$	$A \neq B$

Similarly, given **Poss 4**, *Schema 1* may be transformed into *Schema 3*:

$\neg P(A \sim B)$	$\neg P(AB)$
$P(AB)$	$P(A \sim B)$

Furthermore, with the help of indefinite concepts, one can form *Schema 4*:

$\exists Y(A = YB)$	$\exists Y(A = Y \sim B)$
$\forall Y(A \neq Y - B)$	$\forall Y(A \neq YB)$

Leibniz used to work with various elements of these representations, often combining them into inhomogeneous schemata such as *Schema 5**:

$\exists Y(A = YB)$	$\exists Y(A = Y \sim B)$
$\exists Y \exists Z(YA = ZB)$	$\exists Y \exists Z(YA = Z \sim B)$

But here the representations of PA and PN are inadequate because these formulas are *theorems of L2*! The conditions for PA and PN, however, may easily be corrected by adding the requirement that *YA* is self-consistent. One thus obtains *Schema 5*:

$\exists Y(A = YB)$	$\exists Y(A = Y \sim B)$
$\exists Y \exists Z(P(YA) \wedge YA = ZB)$	$\exists Y \exists Z(P(YA) \wedge YA = Z \sim B)$

3 Primaria Calculi Logici Fundamenta

In this section two drafts of a logical calculus shall be considered which Leibniz had written on one and the same sheet of paper (LH IV, 7B2, 3). The untitled essay on the "recto" side bears the date 1 Aug. 1690; the other essay on the "verso" side is an immediate sequel to the first and bears the title "Primaria Calculi Logici Fundamenta". Both papers have been published in *Cout* (232–235), but only the second one was included in Parkinson's translation (*LLP*, 90–92). Unfortunately, text-critical versions from the Academy-edition are not yet available. The following text contains some deleted passages which were not reproduced in Couturat's nor, therefore, in Parkinson's edition. I did not, however, aim at providing a full apparatus of all variants of the text[19] but included only those passages which seem essential for a full understanding of the problems dealt with

19 I rather leave this task to the professional staff of the Leibniz Research Centre at the University of Münster.

here. These passages have been translated into English, while the original Latin text is reproduced in the footnotes.

3.1 The first draft

> August 1st, 1690
> Each categorical proposition can be conceived as a simple term to which either 'is' or 'is not' is added ("secundi adjecti"). Thus 'Every man is rational' can be conceived as 'Man not-rational is not (or is "not being")'.
> 'Some man is learned' yields 'Man learned is "being"'
> 'No man is a stone' yields 'Man stone is "not being"'
> 'Some man is not learned' yields 'Man not learned is "being"'
> From this conception the laws of conversion and of opposition become immediately evident. Thus, the UN and the PA can be simply converted because in the above reduction both terms are treated in the same way. Nevertheless it is evident that the reduced version differs from the original one, i.e. 'Some man is learned' is different from 'Man learned is "being"' since by the latter it is simultaneously expressed that some man is learned and that some learned is a man.
> There is an opposition between the UN and the PA, namely between 'AB is not "being"' and 'AB is "being"'.
> There is an opposition between the UA and the PN, namely between 'A Not-B is not "being"' and 'A Not-B is being"'.

Leibniz begins his investigations with the homogeneous *Schema 3* in which the laws of opposition are trivially satisfied. Clearly, the formula for the PN, $P(A \sim B)$, is the negation of the UA, $\neg P(A \sim B)$, just as the PA, $P(AB)$, directly negates the UN, $\neg P(AB)$. The next step, however, turns out to be much more difficult to prove:

> But let us see how subalternation or subsumption can be derived from our conception.
> 'Every man is an animal', therefore 'Some man is an animal'. '*A* Not-*B* is not "being"', therefore '*AB* is "being"'.
> 'No man is a stone', therefore 'Some man isn't a stone'. '*AB* is not "being"', therefore '*A* Not-*B* is "being"' and '*B* Not-*A* is 'being".
> The following inference is not valid: '*AB* is "being"', hence '*A* Not-*B* is not "being"'.
> That something is not "being" cannot be inferred in a regular way unless there is a contradiction such as '*A* Not-*A* is not "being"'.

According to *Schema 3*, the laws of subalternation amount to the conditions that $\neg P(A \sim B)$ entails $P(AB)$, and that $\neg P(AB)$ entails $P(A \sim B)$, respectively. In other words, at least one proposition from the pair $P(AB)$, $P(A \sim B)$ must be true. It may well happen that *both* propositions are true, for, as Leibniz rightly remarks, the *truth* of one proposition, say $P(AB)$, doesn't entail that the other, $P(A \sim B)$, has to be false. In general, a negative proposition of the type $\neg P(B)$ can only be proven

by finding a contradictory concept $A \sim A$ contained in B. Anyway, Leibniz goes on as follows:

> This inference must be proved: 'A Not-B is not "being"', hence 'AB is "being"' i.e. this inference must be proved: 'Every A is B', therefore 'Some A is B'.
> I have proved this inference somewhere else as follows: 'Every A is B'. 'Some A is A'; therefore 'Some A is B'. But this proof presupposes a syllogism of the First figure. Namely 'Every A is B'; 'Some C is A', therefore 'Some C is B'. By reduction: 'A Not-B is not "being"', 'AC is "being"', therefore 'CB is "being"'. How can this inference be proved?

Leibniz here reminds himself that the wanted proof of:

Poss 6*: $\neg P(A \sim B) \to P(AB)$.

might be obtained in the same way as, in a previous paper, he had derived the informal law of subalternation from the syllogism *Darii*. On the background of *Schema 3*, *Darii* takes the form

Poss 7: $\neg P(A \sim B) \wedge P(AC) \to P(CB)$.

The "trick" of the syllogistic proof of subalternation consists in simply setting $C = A$[20]. This yields the inference $\neg P(A \sim B) \wedge P(AA) \to P(AB)$, i.e. after an easy transformation:

Poss 6: $P(A) \to (\neg P(A \sim B) \to P(AB))$.

The latter principle shows that Leibnz's earlier subalternation principle **Poss 6*** has to be restricted to self-consistent concepts A. For clearly, if one starts from the impossible concept $A = C \sim C$, then both AB and $A \sim B$ will be self-contradictory, too. Furthermore, a proof of **Poss 6** now becomes easily available, provided that one is willing to assume **Poss 4A**, $A \in B \leftrightarrow \neg P(A \sim B)$, as an axiom. For if $\neg P(A \sim B)$, then $A \in B$ and hence $A = AB$ (according to **Cont 1**); therefore the premise $P(A)$ immediately yields $P(AB)$!

Unfortunately, Leibniz did not discover this proof but continued as follows:

> Since the validity of the inference [of subalternation] is not easily discernible from this reduction [*Schema 3*], it cannot be regarded as the optimal resolution. Thus it is better to reduce everything to equivalences, i.e. to equations:

20 Cf. "Of the Mathematical Determination of Syllogistic Forms" in *LLP*, 107: "Subalternation [...] is proved as follows: Every A is B, Some A is A, therefore Some A is B, which is an argument in Darii."

> $A = YB$ is the UA, where the adjunct Y is like an additional unknown term: 'Every man' is the same as 'A certain animal'.
> $YA = ZB$ is the PA. 'Some man' or 'Man of a certain kind' is the same as 'A certain learned'.
> $A = Y$ not-B. 'No man is a stone', i.e. 'Every man is a not-stone', i.e. 'Man' and 'A certain not-stone' coincide.
> $YA = Z$ not-B. 'A certain man isn't learned' or 'is not-learned', i.e. 'A certain man' and 'A certain not-learned' coincide.

So Leibniz now resorts to the much more complicated and inhomogeneous *Schema 5** and he announces optimistically:

> From these [representations] all principles are demonstrated, for example:
> 'Every man is an animal'; therefore 'Some man is an animal'. For $A = YB$, hence ZA[21] $= ZYB$.
> Set $ZY = W$, then $ZA = WB$.
> 'No man is a stone'; therefore 'Some man is not a stone'; [is proved] in the same way. For $A = Y$ not-B, hence $ZA = ZY$ not-B, i.e. $ZA = W$ not-B.

These elliptic proofs are *formally* correct, for if there exists, e.g., a concept Y such that $(A = YB)$, one can choose some such Y, i.e. set $A = YB$. Hence for arbitrary Z, $ZA = ZYB$; so if one sets $ZY = W$, it follows that $\exists Z \exists W(ZA = WB)$, i.e. the PA according to *Schema 5** is satisfied. However, this demonstration is less worth than it might appear because — as was pointed out in section 2.2 above — the conclusion $\exists Z \exists W(ZA = WB)$ is a tautology![22] Anyway, having thus »proved« the laws of subalternation, Leibniz next turns to the principles of conversion. The conversion of the PA is trivial:

> 'Some man is learned'; therefore 'Some learned is a man'. $YA = ZB$, hence $ZB = YA$.

But the corresponding conversion of the UN reveals a serious difficulty. Let us consider Leibniz's first approach which was afterwards crossed out:

> 'No man is a stone'; therefore 'No stone is a man'. $A = Y$ Not-B, hence Not-A = Not-(YNot-B) = B. This inference presupposes the fundamental equation not- (YNot-B) = B, i.e. negating that some are excluded is to put every.[23]

---

21 As has already been noticed in *Cout* 234, fn. 1, Leibniz erroneously wrote 'ZB' instead of 'ZA'.

22 The same remark applies to the second proof where the formula for the PN, $\exists Z \exists W(ZA = W \sim B)$, is "derived" from the formula for the UN, $\exists Y(A = Y \sim B)$.

23 "Nullus homo est lapis, Ergo Nullus lapis est homo. A = Y non-B, Ergo non-A = non(Y non-B) = B. Supponitur scilicet haec consequentia aequatio fundamentalis non(Y non-B) = B, seu negando quendam excludi est poni omnem." The text continues with the remark "Sed non patet et haec consequentia A = B ergo non-A" then it breaks off.

What Leibniz here temporarily assumes as a "fundamental equation", viz. $\sim(Y \sim B) = B$, can't, however, be correct, because it would entail (by negating both sides of the equation) that $Y \sim B =\sim B$. Hence the premise $A = Y \sim B$ would coincide with $A =\sim B$ while as a matter of fact only $A \in \sim B$ holds by assumption, but not conversely $\sim B \in A$! Furthermore, if one explicates the elliptic equation $\sim(Y \sim B) = B$ by inserting the existential quantifier $\exists Y$, the resulting formula $\exists Y(\sim(Y \sim B) = B)$ becomes a theorem of L2[24], but it doesn't correctly formalize what Leibniz had in mind when he maintained that "negating that some are excluded is to put [or affirm] every". This cryptic remark appears to be an anticipation of the modern law $\neg\exists y\neg By \leftrightarrow \forall yBy$ which establishes a logical relation between the existential and the universal quantifier (ranging over individuals). But such a law cannot, as such, be formalized in L2 since the quantifiers there range over concepts! Anyway, Leibniz must have noticed that something is wrong with the above argument, for he deleted the passage and continued as follows:

> 'No man is a stone', therefore 'No stone is a man' reveals a difficulty in this resolution. Somewhere else we have proved this as follows. 'No man is a stone'. 'Every stone is a stone'. Therefore, 'No stone is a man'. [This is a mood] in the Second Figure, but then the Second Figure must first be proved, although this is not difficult from our [investigations]. Let us first exhibit the difficulty of the proof of the simple conversion of the UN.
> $A = Y$Not-B, hence $B = Z$Not-A. Let us make an analysis. If this inference is valid, then $A = Y$Not-(ZNot-A). Thus it must be shown that A and YNot-(ZNot- A) are the same, e.g., 'man' and 'some not-(some not-man)' coincide, for an arbitrary thing besides man is a certain not-man. Any such thing, for example ZNot-A, can be called W[25]. Then we obtain $A = Y$Not-W. For man, at any rate, is one of those things which are Not-W. Otherwise a certain A would be W, i.e. $XA = TW$ or $XA = TZ$Not-A which is absurd. For if '$A = Y$Not-W' is false, then '$XA = TW$' is true. This inference still has to be confirmed.

The last paragraph is difficult to understand because of three reasons. First, the subject matter itself is logically quite complicated. Second, Leibniz' "elliptic" formalization by means of the indefinite concepts Y, Z, W, X, and T doesn't make entirely clear which one is to be taken as an existential, and which one as a universal quantifier. Third, Leibniz several times switches between an entirely abstract and a more concrete level where concept A is replaced by 'man'. Let's try to reconstruct his thoughts step by step!

---

24 For let Y be $=\sim B$; then $\sim(Y \sim B) =\sim(\sim B \sim B) =\sim\sim B = B$!

25 For the sake of an easier exposition, I have replaced Leibniz's 'M' by 'W'; therefore in the subsequent passage it became necessary to put 'X' for 'W'.

On the background of *Schema 5**, the law of conversion which Leibniz tries to prove here amounts to a variant of the principle of contraposition. For in view of **Cont 1**, the former **Neg 2**, i.e. $(A \in\sim B) \rightarrow (B \in\sim A)$, is equivalent to:

Neg 11: $\exists Y(A = Y \sim B) \rightarrow \exists Z(B = Z \sim A)$.

Since Leibniz is not entirely sure whether this principle is valid, he argues *indirectly*: If **Neg 11** is correct, then — assuming that for some Y (i): $A = Y \sim B$ — there has to exist some Z such that (ii): $B = Z \sim A$. But then '$Z \sim A$' may be substituted for 'B' in (i), so that one obtains the (elliptic) equation (iii): $A = Y \sim (Z \sim A)$. Next he tries to verify (iii) by setting $W = Z \sim A$, thus obtaining (iv): $A = Y \sim W$. Now if (iv) -- or more explicitly (v): $\exists Y(A = Y \sim W)$ — would not be true, in other words, if the UN 'Every A is not-W' according to *Schema 5** would be false, one might infer that the PA, i.e. 'Some A is W' or (vi): $\exists X \exists T(XA = TW)$ is true. Replacing 'W' again by the expression '$Z \sim A$', one would obtain (vii): $\exists X \exists T(XA = TZ \sim A)$, and finally by setting 'V' for 'TZ' one gets (viii): $\exists X \exists T(XA = V \sim A)$ which, as Leibniz maintains, is "absurd".

This complicated proof raises a number of questions. First, is formula (viii) really "absurd", i.e. is it correct to generalize the former principle of consistency **Neg 3**, $A \neq\sim A$, in such a way that, for every Y and Z, YA must be different from $Z \sim A$:

Neg 12*: $\forall Y \forall Z(YA \neq Z \sim A)$?

Second, is Leibniz's inference from (v) to (vi) logically warranted, i.e. does the falsity of the UN according to *Schema 5**, $\neg\exists Y(A = Y \sim B)$, entail the truth of the PA, $\exists Y \exists Z(YA = ZB)$? In a certain way, this question may trivially be affirmed since the formula $\exists Y \exists Z(YA = ZB)$ itself is a theorem of L2.[26] So what is at stake here is rather, whether the corresponding inference still holds when *Schema 5** is amended to *Schema 5*. In other words, how could Leibniz prove the following law

Neg 13: $\neg\exists Y(A = Y \sim B) \rightarrow \exists Y \exists Z(P(YZ) \wedge YA = ZB)$?

We need not enter this difficult question here because Leibniz himself appears to have felt that his »proof« of NEG 11 was not entirely conclusive. He continued the essay with the subsequent considerations which emphasize the fundamental character of the law of contraposition:

26 This proof is not trivial because Schema 5 is not homogeneous!

> 'Every man is an animal', therefore 'Every not-animal is a not-man', $A = YB$ is the same as Not-$B = Z$ Not-A. This inference is fundamental, and the two expressions are equivalent because of the nature of 'every'.
> Thus I assume these principles: $A = B$, therefore Not-A = Not-B, or vice versa; and $A = YB$, therefore Z Not-A = Not-B, i.e. if 'man' coincides with 'a certain animal', namely 'rational animal', then 'not animal' coincides with 'a certain not-man'. For this depends on the transition from individuals to ideas. When I say 'Every man is an animal', then I want this, that the men are found among the animals, i.e. if something is not an animal, it is neither a man.
> On the other hand, when I say 'Every man is an animal', I mean that the notion of animal is contained in the idea of man. And the two methods of approach by notions and by individuals are contrary to each other: Just as all men are part of all animals, i.e. all men are included in all animals, so conversely the notion of animal is in the notion of man; and just as there are other animals besides men, something has to be added to the idea of animal to yield the idea of man; for by augmenting conditions [to the notion], the number [of the individuals] decreases.

The draft ends with the following notes which are eventually crossed out and continued in more detail on the backside of the paper:

> The primary bases of a logical calculus.
> '$A = B$' is the same as '$A = B$ is true';
> '$A \neq B$' is the same as '$A = B$ is false'.
> $A = A$.
> $A \neq B$ Not-A.
> '$A = B$', '$A \neq Not - B$' and '$Not - A = Not - B$' are equivalent.[27]

This deleted passage has been added here because it shows once again that Leibniz was inclined to make the gross mistake of **Neg 9*** where '$A = B$' is equated with '$A \neq\sim B$'.

3.2 The second draft

> The primary bases of a logical calculus.
> (1) '$A = B$' is the same as '$A = B$ is true'
> (2) '$A \neq B$' is the same as '$A = B$ is false'
> (3) $A = A$
> (4) $A \neq B$ Not-A

Instead of (4), Leibniz had originally formulated the weaker principle $A \neq\sim A$, but then he generalized **Neg 3** by adding 'B'. This amounts to the assumption that,

27 "Primaria Logici fundamenta: idem est $A = B$ et $A = B$ est vera. Idem est A non $= B$ et $A = B$ est falsa. $A = A$. Eadem sunt $A = B$ et A non = non B et non A = non B."

for *every* B, A is different from $B \sim A$. In the formalism of *L2* this principle takes the form

Neg 14*: $\forall Y(A \neq Y \sim A)$.

In contrast to **Neg 3**, **Neg 14*** is not unrestrictedly valid but holds only under the assumption that A is self-consistent:

Neg 14: $P(A) \rightarrow \forall Y(A \neq Y \sim A)$.

Take any A such that $P(A)$, and assume, that for some Y, $A = Y \sim A$. Adding 'A' on both sides of this equation yields $AA = AY \sim A$, hence $A = AY \sim A$. But this is absurd because A was assumed to be self-consistent while $AY \sim A$ contains $A \sim A$ and hence is impossible.

Leibniz goes on listing some further fundamental principles:

(5) 'A = Not-Not-A'
(6) $AA = A$.
(7) $AB = BA$.
(8) '$A = B$', 'Not-A = Not-B', 'A not $\neq B$' are the same.

At the end of the last line, Leibniz had originally added '$A \neq$ Not-B'[28] as another formula presumably equivalent to the expressions of (8); but he immediately recognized that this assumption is untenable. As a matter of fact, it would repeat the gross mistake of **Neg 9*** according to which $A = B$ is equivalent to $A \neq\sim B$. But, as Leibniz sets out to explain, only one "half" of this equivalence, namely the inference $A = B \rightarrow A \neq B$, is valid:

(9) If $A = B$, it follows that $A \neq$ Not-B. I prove this in this way. If it does not follow, let $A =$ Not-B (by assuming the contrary). Therefore (by hypothesis) $B =$ Not-B, which is absurd. It can also be proved in this way. $B \neq$ Not-B (by 4), therefore [because of the premise $A = B$] $A \neq$ Not-B.
(10) If $A = AB$, there can be assumed a Y such that $A = YB$. This is a postulate, but it can also be proved, for A itself at any rate can be designated by Y.
(11) If $A = B$, then $AC = BC$. But it does not follow: $AC = BC$, hence $A = B$. For let $A = BC$, then one obtains $AC = BC$ by [11][29] and (6).
(12) '$A = AB$' and 'Not-B = Not-B Not-A' coincide.
(13) If $A = YB$, it follows that $A = AB$. I prove this as follows. $A = YB$ (by hyp.), therefore $AB = YBB$ (by [11]) $= YB$ (by 6) $= A$ (by hyp.). Hence the universal affirmative can be expressed as follows: $A = AB$ or $A = YB$.

28 In the manuscript: 'A non = non B'.
29 Leibniz erroneously has '10' here.

A few comments are in order here.

- Principle (10) exhibits one of the few passages where Leibniz explicitly uses the existential quantifier "there can be assumed a Y such that". The principle itself offers a straightforward application of the fundamental rule of existential generalization, **Exist 1**, since from '$A = AB$' it is inferred that $\exists Y(A = YB)$.
- The converse inference is later proved in (13) so that the UA can *alternatively* be formalized by '$A = AB$' or by '$\exists Y(A = YB)$'.
- Principle (11) not only asserts the law of "addition", **Iden 6**, but also states that the converse principle of "subtraction", $AC = BC \rightarrow A = B$, is invalid. Leibniz's proof is a bit elliptic but it can be completed as follows: Let $A = BC$ in such a way that C is an "essential" component of "A", i.e. such that A does not coincide with B alone. Then the premise $A = BC$ entails $AC = BCC = BC$; but in *this* equation, $AC = BC$, one may not "subtract" concept C since by assumption $A \neq B$.
- Principle (12) represents a version of the law of contraposition, **Neg 2**, where the $\in$-expressions are replaced by equations according to **Cont 1** .

Next Leibniz considers several other formalizations of the remaining categorical forms:

> The particular affirmative thus: $YA = YAB$ or $YA = ZB$, or also $AB = AB$, i.e. AB is "being" or [AB and AB] can stand for each other, or $A \neq A$ Not-B.
> The universal negative 'No A is B' thus: $A = Y$ Not-B, *i.e.* $A = A$ Not-B, *i.e.* AB is "not being".
> The particular negative 'Some A isn't B' thus: $A \neq AB$, or A Not-B is "being".

It is somewhat surprising to see that Leibniz considers no less than five different formalizations for the PA. The fourth and fifth condition, '$P(AB)$' and '$A \neq A \sim B$', are familiar from *Schema 3* and *Schema 2*, respectively. The first two expressions, '$YA = YAB$' and '$YA = ZB$', are based on the idea of abbreviating 'Some A' by 'YA' and thus to represent the PA by '$YA \in B$'. This formula can be transformed by means of **Cont 1** into '$YA = YAB$' or by means of **Cont 4** into '$YA = ZB$'. But both conditions are *inapt* to represent the PA, because, as was stressed several times before, they are mere *tautologies*![30]

The same holds for the third condition '$AB = AB$' which, according to **Iden 1**, is a primitive tautology of *L1*! The background for this unusual approach is an idea already discussed in the *GI*. In § 128 Leibniz argues that if a PN 'No A is B'

30 To be somewhat more precise, the corresponding formulas with added quantifiers, $\exists Y(YA = YAB)$ and $\exists Y \exists Z(YA = ZB)$, are theorems of *L2*.

is true, then $A = A \sim B$ and hence (according to **Poss 4a**) "AB is an impossible, or rather a false term". Leibniz doubts whether such a "false" term may ever be consistently used within logical inferences. In particular, he is not certain whether the fundamental identity '$A = A$' holds for such a "false" term. Therefore affirming the identity '$AB = AB$' (or affirming that 'AB' and 'AB' can "stand for each other") appears to be tantamount to saying that 'AB' is not a "false", but a self-consistent ("being") term. Consequently, in § 152 he suggests to formalize the PA alternatively by '$AB = AB$' or by '$P(AB)$'. But a bit later (§ 155) he concludes that "all things considered, then, it will perhaps be better for us to say that, in symbols at least, we can always put $A = A$, though nothing is usefully concluded from this when A is not a thing".

The three formalizations of the UN are all correct. In view of the proof given in (10), (13), '$A = Y \sim B$', or more explicitly '$\exists Y(A = Y \sim B)$', is provably equivalent to the standard requirement according to *Schema 2*, '$A = A \sim B$'; and the other condition '$\neg P(AB)$' is familiar from *Schema 3*.[31] Finally, the PA can adequately be represented either by '$A \neq AB$' according to *Schema 2* or by '$P(A \sim B)$' according to *Schema 3*. In what follows, however, Leibniz decides to continue with the simpler *Schema 2*:

> But let us see if the following alone are sufficient:
> Univ. Aff. $A = AB$. Part. Neg. $A \neq AB$. Univ. Neg. $A = A$ Not-B. Part. Aff. $A \neq A$ Not-B.
> If $A = AB$, then $A \neq A$ Not-B, i.e. Part. Aff. follows from Un. Aff.
> Proof: Let $A = A$ Not-B (by assuming the contrary). Since $A = AB$ (by hyp.) we obtain A Not-$B = AB$, which is absurd by (4). Or more briefly: A Not-$B \neq AB$ (by 4); if one substitutes here 'A' for 'AB' (for they are equivalent by hyp.) one gets A Not-$B \neq A$, Q.E.D.
> If $A = A$ Not-B, then $A \neq AB$, i.e. Part. Neg. follows from Univ. Neg.
> Proof: A not-$B \neq AB$ (by 4). Substitute 'A' for 'A not-B' (for they are equivalent, by hyp.), and one gets $A \neq AB$.
> '$A \neq A$ Not-B' and '$B \neq B$ Not-A' are equivalent, i.e. the particular affirmative proposition can be simply converted.

So here again, like in the previous draft, Leibniz faces the problem of having to prove the law of *conversion* of the PA which, in view of *Schema 3*, now takes the shape of the following variant of the law of *contraposition*:

Neg 15: $A \neq A \sim B \leftrightarrow B \neq B \sim A$.

31 It may be worthwhile mentioning that in a preliminary version of the text Leibniz had erroneously mixed up the negation in the corresponding formulas: "Universalis negativa hic: non $A = YB$ vel non $A =$ non AB".

A first attempt (afterwards deleted) runs as follows:

> Proof: Set (1) $A \neq A \sim B$ (by hyp.), then I say B will be $\neq B$ Not-A. For if this would not be the case, then (2) '$B = B$ Not-A' would be true (by the contrary hyp.), and if this value of B is substituted in (1) one obtains $A \neq A$ Not-(B Not- A) what is absurd because $A = A$ Not-(B Not-A).[32]

This argument is *formally correct*, and it could be accepted as a proof of **Neg 15** if the concluding assertion '$A = A \sim (B \sim A)$' might be taken for granted. As a matter of fact, this formula is a *theorem* of *L1*,

Neg 16: $A = A \sim (B \sim A)$.

In view of **Cont 1**, **Neg 16** says as much as $A \in\sim (B \sim A)$, which can be derived in *L1* as follows: Since, by the trivial **Conj 4**, $B \sim A \in\sim A$, the ordinary law of contraposition, **Neg 2**, entails $\sim\sim A \in\sim (B \sim A)$ and hence, by eliminating double negation, $A \in\sim (B \sim A)$. However, this consideration doesn't represent a real proof of **Neg 16** (nor, therefore, a proof of **Neg 15**), but only a *derivation* of *one* variant of the law of contraposition from the *other*.[33]

Leibniz apparently was not satisfied with this result for he started a second attempt:[34]

> Proof: From $A \neq A$ Not-B it follows (by 9) $B \neq B$ Not-A. Therefore also conversely; or immediately, '$A = A$ Not-B' coincides with '$B = B$ Not-A' (by 9), therefore also their negations coincide. Q.E.D.

The problem with this "proof", however, is that principle (9), which Leibniz here relies on, is (much) too weak to warrant the crucial inferences. Leibniz must have felt this, for he makes a third attempt.

---

32 "Demonstratio: Ponatur (1) A non $= A$ non B (ex hyp.) ajo fore B non $= B$ non A. Sit enim falsum si fieri potest, ergo verum erit (2) $B = B$ non A (ex hyp. contraria) qui valor ipsius B substituatur in 1 fiet A non $= A$ non (B non A) quod est absurdum nam $A = A$ non (B non A)".

33 Conversely, **Neg 2** may be derived from **Neg 16** as follows. Assume $\sim B \in\sim A$, hence $\sim B =\sim B \sim A$ (by **Cont 1**) and therefore (by **Iden 5** and **Neg 1**) $B =\sim (\sim B \sim A)$; hence from **Neg 16** (with '$\sim B$' substituted for 'B'), i.e. $A = A \sim (\sim B \sim A)$, one gets $A = AB$, i.e. $A \in B$. Hence we have shown ($\sim B \in\sim A \rightarrow A \in B$), which is just one "half" of **Neg 2**; the second "half" follows in the same way.

34 The following text was preceded by a deleted passage (not edited in *Cout*) which contains the same idea: "Idem sic demonstrari potest A non $= A$ non B (ex hyp.) Ergo (per 9) B non $= B$ non A."

> Let us see if we can deduce 'B = B Not-A' from 'A = A Not-B' in another way. If $A = A$ Not-B, then $AB = AB$ Not-B, hence AB is "not being". But if we deduce '$A = A$ Not-B' from 'AB' is "not being", we might with equal justice deduce its reciprocal '$B = B$ Not-A'.

This argument has to be reconstructed as follows: Since $A = A \sim B$ entails $AB = AB \sim B$, it follows that $\neg P(AB)$. Hence, as a variant of principle **Poss 4a** one obtains

Poss 8a: $A = A \sim B \rightarrow \neg P(AB)$.

Now $\neg P(AB)$ is equivalent to $\neg P(BA)$; therefore – Leibniz appears to argue — one may conversely infer the reciprocal formula $B = B \sim A$ from $\neg P(BA)$. But this argument overlooks that, unlike **Poss 4a**, **Poss 8a** has only been proven as a (one-way) *implication* but not as an *equivalence*! Again Leibniz must have felt that his »proof« was not conclusive, for he made a fourth attempt:

> Perhaps it can be proved as follows without making any supposition. Let AB be "being", then $A \neq A$ Not-B; for if A would be = A Not-B, AB would be = AB Not-B and so AB would not be "being" which is contrary to the hypothesis. With equal justice [it follows] $B \neq B$ Not-A. When it is said that AB is "being" or not "being", it is presupposed that A and B are "being". Let us see if it can be shown conversely: $A \neq A$ Not-B, therefore AB is "being"! Now if, assuming A and B to be "being", AB were not "being", then one of them, A or B, would evidently involve the contradictory of what the other involves. Let us assume, therefore, that A involves C and B involves Not-C (from which, again, it follows that B involves D and A involves Not-D, namely D = Not-C). Let $A = EC$ and $B = F$ Not-C. Now $EC = EC$ Not-(F Not-C)[35], i.e. EC contains Not-(F Not-C) (or, whatever involves C, involves the negation of that which negates C). That is, $A = A$ Not-B which is contrary to the hypothesis.
> Therefore 'AB is "being"', '$A \neq A$ Not-B' and '$B \neq B$ Not-A' are equivalent, i.e. follow from each other mutually.
> Similarly, 'AB is not "being"', '$A = A$ Not-B' and $B = B$ Not-A' are equivalent.
> So we have found the key which permits us to use the reduction of complex to incomplex terms.

As the concluding sentence shows, Leibniz believes to have eventually found a proof of the fundamental relation between "complex terms", i.e. propositions of type $A \in B$ (or $A \in\sim B$), and "incomplex terms", i.e. assertions like '$A \sim B$ is not "being"' (or 'A not-B is not "being"'). This fundamental relation, which in the *GI* had been formulated, e.g., by **Poss 4**, now takes the shape

Poss 8a: $A = A \sim B \leftrightarrow \neg P(AB)$.

35 This important principle (see **Neg 17** below) had been underlined by Leibniz; this emphasis obviously escaped Couturat's attention.

If this principle really had been *proved*, then also the crucial principle **Neg 15**, which was intrinsically at stake all the time, could be considered as proven. However, a closer inspection of the foregoing text reveals that only one "half" of the equivalence, namely $A = A \sim B \rightarrow \neg P(AB)$, actually was proved[36], while the "proof" of the converse implication, $\neg P(AB) \rightarrow A = A \sim B$, rests on two additional (unproven!) principles, namely:

Poss 9: $P(A) \wedge P(B) \wedge \neg P(AB) \leftrightarrow \exists Y(A \in Y \wedge B \in\sim Y)$.

Neg 17: $EC = EC \sim (F \sim C)$.

Now, **Poss 9** turns out to be deductively equivalent to **Poss 8b** [37], and **Neg 17** is easily shown to be deductively equivalent to **Neg 16.**[38] Hence we end up with the following situation.

At the beginning of the "Primaria Calculi Logici Fundamenta", Leibniz listed a number of *fundamental* principles, i.e. *axioms* which need not themselves be *proved* but which may rightly be assumed as a basis for proving other theorems. Among these "Fundamenta" he had originally subsumed the principle of contraposition in the form of (12), $A = AB \leftrightarrow\sim B =\sim B \sim A$. When Leibniz later tries to derive the laws of the theory of the syllogism from the "Fundamenta", he is thrown back to the principle of contraposition. But now he is no longer willing to use either (12) or its counterpart **Neg 2**, $A \in B \leftrightarrow\sim B \in\sim A$, as an axiom, but obstinately tries to prove it. In the end, however, Leibniz at best derives (12) from other principles like

Neg 15: $A \neq A \sim B \leftrightarrow B \neq B \sim A$

---

36 This proof is exactly the same argument that had been analyzed above in connection with **Poss 8a**.

37 Note, first, that Leibniz's premises $P(A)$ and $P(B)$ are redundant. If $\neg P(A)$, then A contains every concept, hence in particular $A \in\sim B$ so that there exists a Y, viz. $Y =\sim B$, such that $A \in Y \wedge B \in\sim Y$. Similarly, if $\neg P(B)$, B contains every concept, hence $B \in\sim A$, so that again there exists a Y, viz. $Y = A$, such that $A \in Y \wedge B \in\sim Y$. Hence **Poss 9** may be simplified to $\neg P(AB) \rightarrow \exists Y(A \in Y \wedge B \in\sim Y)$. This principle is deductively equivalent to the crucial »half« of **Poss 8b**, $\neg P(AB) \rightarrow A = A \sim B$, i.e. $\neg P(AB) \rightarrow A \in\sim B$, because (i) if there exists a Y such that $A \in Y \wedge B \in\sim Y$, then $\sim\sim Y \in\sim B$ and hence $A \in\sim B$; (ii) if conversely $A \in\sim B$, then trivially there exists a Y such that $A \in Y \wedge B \in\sim Y$, namely $Y =\sim B$!

38 Since 'E' may be chosen in **Neg 18** as an arbitrary concept, one may in particular set $E = C$ so that one obtains: $CC = CC \sim (F \sim C)$, i.e. $C = C \sim (F \sim C)$, i.e. an alphabetic variant of **Neg 16.** Conversely, **Neg 18** is immediately obtained from **Neg 16**, i.e. $C = C \sim (F \sim C)$, by the "addition" of 'E'.

Neg 16: $A = A \sim (B \sim A)$
Neg 17: $CA = CA \sim (B \sim A)$

which constitute just variants of the same law. Or he derives the principle of contraposition from principles like **Poss 8**, **Poss 9**, which are in a certain sense even stronger than **Neg 2** and which should therefore also not be accepted as "fundamental".

The draft of the "Fundamenta" ends with some miscellaneous considerations:

> In each term A or not-A is contained, if A is not contained, not-A will be contained, and the other way round; therefore 'not containing A' and 'containing not-A' are equivalent, or '$A = Y$ not-B' and '$A \neq ZB$' are equivalent, or '$A = A$ not-B' and '$A \neq AB$' are equivalent. Hence bad.[39]
> Hence we have arranged matters better in the following paper of 2 August 1690.
> Not-AB is contained in not-B, or not-B = not-B not-AB.
> If $A = BC$, is $A : C = B$ where this is to be understood that C is removed from A? Reducing this to primitive terms, let $B = CE$, then $A = CEC$ or $A = CE$, so that $A : C$ is not always $= B$. So this is only valid in the case of primitive terms.
> Wherever we generally have EB and 'E' is understood as 'any', 'B' can be substituted [for 'EB']; for taking ['B'] for '[E]'[40], one obtains $EB = BB = B$.
> If not-$AB \neq A$ not-B, then not-$AB = B$ not-A, and vice versa, i.e. 'Not-$AB \neq A$ not-B' and 'Not-$AB = B$ not-A' are equivalent.

The first paragraph nicely illustrates one of the main problems of Leibniz's logic – the *lack of the operator of conceptual disjunction*! To be sure, the operator '$A \cup B$' might easily be introduced in *L1* by *definition*, namely as the negation of the conjunction of the negated disjuncts:

Disj 1 $A \cup B :=\sim (\sim A \cap \sim B)$

Furthermore, Leibniz appears to have known this "De-Morgan-law", at least as a law for the corresponding *propositional* operators, $\alpha \vee \beta \leftrightarrow \neg(\neg\alpha \wedge \neg\beta)$.[41] But within all his drafts of a universal calculus, Leibniz never seriously took conceptual

39 This passage is missing in *LLP*.

40 Leibniz erroneously has 'sumendo E pro B' instead of 'sumendo B pro E'.

41 Cf. *Acad* VI, 4, 899: "'Or' is the negation of a negative pair; it is worthwhile noting that while the negation of a negation is an affirmation, the negation of a negative pair is not a simple affirmation but an alternative affirmation; thus 'Peter or Paul is coming' is the same as to say 'It is false that neither Peter nor Paul is coming'. This proposition can also be reduced to the two following propositions: 'If Peter is not coming, then Paul is coming' and 'If Paul is not coming, then Peter is coming'."

conjunction into account. When he now notes that "in each term A or not-A is contained", he may somehow have "felt" that the tautological concept 'A or not-A' is contained in every concept B:

Disj 1 $B \in A \cup \sim A$.

But since he doesn't have the appropriate tool of *conceptual* disjunction at hand, he tries to paraphrase this by means of *propositional* disjunction as

Neg 18* $B \in A \vee B \in \sim A$.

This, however, is just a variant of the invalid "half" of the notorious principle **Neg 8***. Leibniz gradually recognizes the invalidity of **Neg 18*** when he first transforms it into the assertion that '$B \notin A$' and '$B \in A$' are equivalent, and then further into the statements that '$A \neq A \sim B$' and '$A = AB$' are equivalent, or (according to **Cont 4**) that '$\exists Y(A = Y \sim B)$' and '$\forall Z(A \neq ZB)$' are equivalent.[42]

Furthermore Leibniz mentions the law $\sim B = \sim B \sim (AB)$, i.e. $\sim B \in \sim (AB)$, which is easily shown to be deductively equivalent to the principle of contraposition, **Neg 2**.[43]

Next he remarks that one may not simply "subtract" a conjunct C from the "sum" BC. E.g., if $B = CE$ and $A = BC$, so that $BC = CEC = CE$, it doesn't follow that $B = E$. The hint that such a "subtraction" is valid only in the case of primitive terms most likely refers to the calculus of "real addition" which Leibniz had developed above all in the paper "Non inelegans specimen demonstrandi in abstractis" of around 1687.[44]

The next brief remark apparently has to be understood as follows: If a formula α contains 'EB' where the indefinite concept E functions as a universal quantifier ("any B"), then $\alpha[EB]$ entails $\alpha[B]$. This law of *L2*,

Univ 1 $\forall Y \alpha[YB] \rightarrow \alpha[B]$,

is rather trivial and it doesn't seem to stand in any relevant relation to the other problems and principles under discussion here.

---

42 It should be noted that during the last two steps of transformation, Leibniz interchanges the variables 'A' and 'B'.

43 Assuming **Neg 2**, the trivial $AB \in B$ immediately yields $\sim B \in \sim (AB)$. Conversely **Neg 2** can be derived as follows: Suppose $A \in B$; then $A = AB$ and further $\sim A = \sim (AB)$; hence in $\sim B \in \sim (AB)$ one may substitute '$\sim A$' for '$\sim (AB)$' thus obtaining the desired conclusion $\sim B \in \sim A$.

44 Cf. *Acad* VI 4, 845–855; a discussion of this calculus may be found in Lenzen (2000).

The final observation, however, once again appears to be related to the notorious problem of the lack of conceptual disjunction. Leibniz somehow "feels" that the negation of AB is equivalent to the disjunction of $\sim A$ and $\sim B$:

Disj 3 $\sim(AB) = (\sim A) \cup (\sim B)$.

But since he doesn't have '$\cup$' as a conceptual operator at hand, he tries to paraphrase this law as the *implication*

Neg 19a* $\sim(AB) \neq A \sim B \rightarrow \sim(AB) = B \sim A$.

which might as well be rephrased as the disjunction

Neg 19b* $\sim(AB) = A \sim B \vee \sim(AB) = B \sim A$.

Leibniz further recognizes that (because of the symmetry $AB = BA$) "vice versa" $\sim(AB) \neq B \sim A$ would then entail $\sim(AB) \neq A \sim B$, so that **Neg 19a*** might be strengthened into the equivalence

Neg 19a* $\sim(AB) \neq A \sim B \leftrightarrow \sim(AB) = B \sim A$.

However, none of these variants is valid! Consider, e.g., the special case $B = \sim A$. Then $\sim(AB) = \sim(A \sim A)$ becomes the tautological concepts, $\top$, and **Neg 19b*** therefore maintains, that $\top$ either coincides with $A \sim B = A \sim\sim A = AA = A$, or $\top$ coincides with $B \sim A = \sim A \sim A = \sim A$. As a matter of fact, however, there are many concepts A such that neither A nor $\sim A$ is a tautology!

4 Fundamenta Calculi Logici

August 2^{nd}, 1690
The Bases of the Logical Calculus
(1) '$A = B$' is the same as '$A = B$ is a true proposition'.
(2) '$A \neq B$' is the same as '$A = B$ is a false proposition'.
(3) $A = AA$, i.e. the addition of a letter to itself is here redundant.
(4) $AB = BA$, i.e. transposition makes no difference.
(5) '$A = B$' means that one can be substituted for the other, 'B' for 'A' and 'A' for 'B'; i.e. they are equivalent.

The following two paragraphs deleted by Leibniz are missing in Parkinson's edition although they had been edited by Couturat:

> (6) What contains *A* not-*A* is "not being" or a "false term".
> (7) Every term contains *A* or not-*A*.

Principle (6) will be picked up in (9) below. (7) might *theoretically* be interpreted as the correct principle **Disj 2** according to which every *B* contains the tautological concept '*A* or not-*A*'. But there is little evidence that Leibniz had this interpretation in mind. In all likelihood, what he meant is rather the *wrong* principle that, for every *B*, either *B* contains *A*, or *B* contains not-*A*. As has been explained already in section 2.1 above, the inference from $B \notin A$ to $B \in\sim A$ holds only in the rare special case where B is an individual concept! The fact that Leibniz crossed out (7) indicates that he recognized the invalidity of this principle. Anyway he continued as follows:

> (6) 'not' immediately repeated cancels itself.
> (7) Therefore *A* = Not-Not-*A*.
> (8) Further, '*A* = *B*' and '*A* not ≠ *B*' are equivalent.
> (9) What contains '*A* not-*A*' is "not being" or a false term; e.g., if *C* = *AB* Not-*B*, *C* would be "not being".

Principle (9) is just a variant of **Poss 1**, i.e. it defines that a concept *C* is impossible if and only if $C \in B \sim B$. In the margin Leibniz adds a grammatically awkward remark probably to be connected to (9):

> A false proposition results if, by the admittance of it, terms which are assumed as true yield something false.

The following deleted paragraph[45] was not edited by Couturat:

> (10) In every term *A* or not-*A* is contained, i.e. if $B \notin CA$ (*breaks off*).

Again it is *theoretically* possible that Leibniz here had envisaged the correct principle **Disj 2**: $B \in A \cup\sim A$, but in all likelihood he only repeated the old mistake of assuming **Neg 18***, i.e. $(B \in A \vee B \in\sim A)$, or $B \notin A \rightarrow B \in\sim A$. For he paraphrased '$B \notin A$' by 'if $B \neq CA$' before recognizing that it would be wrong to continue 'then *B* = *D* not-*A*', i.e. $B \in\sim A$.

45 (10) In omni termino inest *A* vel non *A*, seu si *B* non = *CA* (*breaks off*).

The following three articles are unproblematic.

(10) '$A \neq B$' and '$B \neq A$' are equivalent. This follows from (5).
(11) '$A = B$' and 'not-A = not-B' are equivalent, for since A can be substituted for B by (5), so substitution in 'not-A' yields 'not-B', i.e. for 'not-B' can be substituted 'not-A'. Similarly it is shown that 'not-A'[46] can be substituted for 'not-B'. Hence, because A and B can be substituted for each other, i.e. since $A = B$, also 'not-A' and 'not-B' can be substituted for each other, i.e. not-A = not-B. But as not-A = not-B has been derived from $A = B$, so also not-not-A = not-not-B, i.e. $A = B$, will be derived from not-A = not-B. Hence these truths follow from each other or are equivalent.
(12) If $A = B$, then $AC = BC$. This is proved from (5). But it does not follow: $AC = BC$, therefore $A = B$. For if A would be only = BC, then one would obtain (by 3) $AC = BC$.

(12) expresses the trivial principle **Iden 6** together with the observation that the converse implication is not valid. As was explained already in section 3.2, one may not simply "subtract" a concept C from the equation $AC = BC$. The above proof obviously has to be understood as follows: If A is "only" the same as BC, i.e. if A does not coincide with either B or C alone, then $A = BC$ nevertheless entails $AC = BCC = BC$. Thus if subtracting C would be generally admissible, one would obtain $A = B$ in contradiction to the assumption that A coincides "only" with BC.

The following paragraph deals with variants of the principle of consistency. The manuscript reveals that Leibniz was searching, so-to-speak, for the »strongest possible generalization« of **Neg 3** (or its counterpart **Neg 4**). Leibniz started with the simple **Neg 4**:

(13_1) If $A = B$, then $A \neq$ not-B, otherwise.[47].

This version breaks off and is immediately replaced **Neg 3**:

(13_2) $A \neq\sim A$[48].

Next he strengthens this into:

(13_3) $AB \neq C$ non-B; and therefore also $A \neq\sim A$; and in the same way (leaving out).[49].

46 As was already noted in *Cout* 421, fn. 4, Leibniz erroneously has 'A' instead of 'not-A'.
47 (12) "Si $A = B$ sequitur A non = non B, alioqui"; this was not edited by Couturat.
48 In the original: A non = non A. This principle and the following variant had been numbered as '(12)' while the (final) (12) originally was '(13)'.
49 In the original: AB non = C non-B; atque ideo et A non = non A et eodem modo (omissis).

This version is (partially) crossed out and modified as follows:

> (13_4) $B \neq\sim B$ and even more generally $AB \neq C$ not-B. Proof. For let be (1) $AB = C$ not-B; but (2) $AB = ABAB$ (by art. 3) and $ABAB = ABC$ not-B (by no. 1 of this art.). Therefore, arguing from the first to the last, $AB = ABC$ not-B which is absurd according to (9), for AB would be a false term, i.e. implying a contradiction.

Finally Leibniz inserts the concept letter 'E' in all subformulas 'C non-B' and thus obtains the "official" version:

> (13_5) $B \neq\sim B$ and even more generally $AB \neq C$ not-EB. Proof. For let be (1) $AB = C$ not-EB; but (2) $AB = ABAB$ (by art. 3) and $ABAB = ABC$ not-EB (by no. 1 of this art.). Therefore, arguing from the first to the last, $AB = ABC$ not-EB which is absurd according to (9), for AB would be a false term, i.e. implying a contradiction.[50]

It has already been shown in section 3.2 that **Neg 3** may be generalized by adding a concept C on the "right hand" side[51] of the inequality $B \neq\sim B$. Somewhat more exactly, if B is *self-consistent*, then B can't coincide with $C \sim B$, for any C:

Neg 14 $P(B) \rightarrow B \neq C \sim B$.

Furthermore, in section 3.1 we already encountered the problem whether **Neg 12*** is valid, i.e. whether (or under which premises) concepts A and C may be added on both sides of the inequality '$B \neq\sim B$' to yield '$AB \neq C \sim B$'. Leibniz himself now provides an answer. As the proof of version (13_4) shows, $AB \neq C \sim B$, provided that 'AB' is not "a false term, i.e. not implying a contradiction":

Neg 12 $P(AB) \rightarrow AB \neq C \sim B$.

Clearly, if $AB = C \sim B$, then adding 'AB' on both sides of the equation yields $ABAB = ABC \sim B$, hence $AB \in B \sim B$, i.e. $\neg P(AB)$.

However, the corresponding argument of version (13_5) doesn't represent a proof for the *stronger* principle:

Neg 20* $P(AB) \rightarrow AB \neq C \sim (EB)$.

---

50 Cf. *Cout* 422; in Couturat's edition the words 'et eodem modo (omissis)' were retained because they had (erroneously) not been crossed out by Leibniz.

51 Of course, every inequality $\delta \neq \epsilon$ is *symmetric* and may therefore be transformed into $\epsilon \neq \delta$. The "right hand side" of the inequality $B \neq\sim B$ here simply means the side which contains the negation operator. If in contrast C is added on the "left hand side" (not containing ~), then the requisite premise is not $P(B)$ but $P(\sim B)$!

where a third concept, E, is inserted within the scope of the negation operator. If A is compatible with B, it doesn't follow that, for arbitrary concepts C and E, $AB \neq C \sim (EB)$. Contrary to Leibniz's claim, '$ABC \sim (EB)$' is *not generally* a "false" term! What can only be maintained is that for any E *compatible with* AB, AB must be different from $C \sim (EB)$:

Neg 20 $P(ABE) \rightarrow AB \neq C \sim (EB)$.[52]

The manuscript continues with some preliminary versions of (14) which have not been edited by Couturat:

> (14_1) $AB = AB$ not-$(C$ not-$B)$
> (14_2)[53] '$A = B$' and '$A \neq$ not-B' are equivalent. For let be $A = B$, then I say it follows $A \neq$ not-B. For if (1) A would be $=$ A not-B, one would obtain $B =$ not-B (by (1)) what is absurd by 13. Similarly, let A not be $=$ not-B, then I say it follows $A = B$. For assume $A \neq$ (*breaks off*).
> (14_3) '$A = A$ not-B' and '$A \neq AB$' are equivalent. For let us first show that '$A = A$ not-B' entails '$A \neq AB$'. For let be (1) $A = A$ not-B; then I say it follows that $A \neq AB$. For if we set, if this is possible, (2) $A = AB$, then we get (from (1)) $A = AB$ not-B what is absurd by (9), or we get (from (2)) $AB = A$ not-B in contradiction to (13). Let us also show that $A = A$ not-B follows from $A \neq AB$. For if we set $A \neq A$ not-B and $A \neq AB$ (*breaks off*).[54]

Somewhat surprisingly version 14_1 doesn't deal with an inequality like those in **Neg 12** and **Neg 20**, but instead puts forward the equality $AB = AB \sim (C \sim B)$ which is just another version of the principle of contraposition:

Neg 16 $B = B \sim (C \sim B)$.

---

52 Since C is an arbitrary concept, the exact meaning of textbfNeg 20 is better brought out by means of a universal quantifier as $P(ABE) \rightarrow \forall Y(AB \neq Y \sim (EB))$. This principle may be proved as follows. Assume (for the sake of reductio ad absurdum) that for some Y, say $Y = C$, $AB = C \sim (EB)$; "adding" EB on both sides yields $ABBE = CEB \sim (EB)$, hence $ABBE = ABE$ contains the contradictory concept $EB \sim (EB)$, i.e. $\neg P(ABE)$. The converse implication $\forall Y(AB \neq Y \sim (EB)) \rightarrow P(ABE)$ also is provable; setting $Y = AB$, $\forall Y(AB \neq Y \sim (EB))$ entails $AB \neq AB \sim (EB)$, i.e. $AB \notin \sim (EB)$; a fortiori $AB \notin \sim E$ (since $\sim E \in \sim (EB)$ by **Conj 4** and contraposition); thus by means of **Poss 4a** one obtains $P(ABE)$.

53 Although in the manuscript (14_2) is standing on top of (14_1), it seems that it was composed later. For (14_2) was squeezed in small letters in an empty space between (14_1) and (13).

54 (14)Aequivalent $A = B$ et A non $=$ non B. Sit enim (1) $A = B$, ajo sequi A non $=$ non B. Nam si esset $A =$ non B foret $B =$ non B (per 1) quod est abs. per art. 13. Similiter sit A non $=$ non B, ajo sequi $A = B$. Est enim A non $=$ (*breaks off*).
$AB = AB$ not-$(C$ not-$B)$
Aequivalent $A = A$ non B et A non $= AB$. Nam ostendamus primo ex $A = A$ non B sequi A non $= AB$. Esto enim (1) $A = A$ non B ajo sequi A non $= AB$. Nam ponamus si fieri potest esse (2) $A = AB$, fiet (ex 1) $A = AB$ non-B quod est absurdum per 9 vel fiet (ex 1) $AB = A$ non B contra 13. Ostendemus et ex A non $= AB$ sequi $A = A$ non B. Nam ponamus A non $= A$ non B et A non $= AB$ (*breaks off*).

Version 14_2 once again contains the gross mistake **Neg 9*** of assuming that $A = B$ would be equivalent to $A \neq\sim B$. Leibniz easily derives $A \neq\sim B$ from $A = B$, but when he tries to construct a similar proof for the converse implication, $A \neq\sim B \rightarrow A = B$, he notices the error and breaks off.

Version 14_3 in a similar way repeats the mistake **Neg 8*** of claiming that $A = A \sim B$, i.e. $A \in\sim B$, is equivalent to $A \neq AB$, i.e. $A \notin B$. Given the tacit assumption that A is self-consistent, Leibniz's proof for the implication $A = A \sim B \rightarrow A \neq AB$ is absolutely correct; but when he tries to construct a similar proof for the converse implication, $A \neq AB \rightarrow A = A \sim B$, he once more notices the error and breaks off.

The final version of (14) runs as follows:

> (14) If $A = B$, it follows that $EA \neq C$ not-FB. For $EA \neq C$ not-FA (by 13); therefore, substituting (by hyp.) 'B' for the last 'A', $EA \neq C$ not-FB. It makes no difference when some proposition is negated.

This proof is formally correct, but it heavily relies on principle (13) which, as was shown above, is valid only in the case where concept F is compatible with EA: $P(EAF)$! A similar critique applies to the subsequent paragraph:

> (15) If $A = FB$, it follows that $EA \neq C$ not-FGB. For $EA \neq C$ not-GA (by 13). Therefore substituting 'FB' for 'A' yields $EA \neq C$ not-FGB.

It remains a bit mysterious why Leibniz dealt with the complicated issue of generalized principles of consistency at all in such a great detail. Probably he was hoping to make use of them in the still unfinished business of proving the principle of *contraposition*. This topic will be picked up in following paragraph 17. First, however, in (16) Leibniz presents another proof of the simple principle of consistency **Neg 5***:

> (16) If $A = A$ not-B, then $A \neq AB$. For $A \neq AB$ not-B (by 9), therefore (substituting 'A not-B' for 'A' by the hyp. here) A not-$B \neq AB$ not-B, therefore $A \neq AB$.

Couturat thought that Leibniz here committed the fallacy of "subtracting" 'C' from $AC = BC$ to infer $A = B$.[55] As a matter of fact, however, what is at stake is not an *equation* but an *inequality*, and it is logically absolutely correct to "subtract" the concept $\sim B$ from '$A \sim B \neq AB \sim B$' to infer '$A \neq AB$'.

The following paragraph crossed out by Leibniz was not edited by Couturat:

> "Scholion: In every term a given term is contained either affirmatively or negatively, e.g. Every" (*breaks off*).

55 Cf. *Cout* 422, fn. 2: "Ici Leibniz paraît conclure de $AB = BC$ à $A = B$, ce qui n'est pas possible en général…".

Here Leibniz considers once again the principle ($A \in B \vee A \in\sim B$) which holds only for individual concepts, but not for arbitrary concepts. Since Leibniz crossed out the sentence, he obviously recognized that the principle is invalid. The paper concludes with the following four paragraphs:

> (17) Not-B = not-B not-(AB), i.e. not-B contains not-AB, or not-B is not-AB.
> This remains to be proved in our calculus.
> (18) $C = C$ not-(A not-C); follows from (17) by putting 'not-C' for 'B'.
> (19) '$A = AB$' and 'not-B = not-B not-A' are equivalent. This is the conversion by contraposition.
> For if (1) $A = AB$ while (2) not-B = not-B not-(AB) (by 17), then putting 'A' for 'AB' in num. 2 yields not-B = not-B not-A. The other way round, if (1) not-B = not-B not-A, while (2) not-B = not-B not-(AB) (by 17), joining (1) and (2) yields $A = AB$. (However, this inference is quite dubious by the note to 12. For we have, though, B not-$A = B$ not-AB, but does this entail $A = AB$? Certainly, if $BC = BD$, then indeed $C = D$ if C and B have nothing in common.)
> (20) Not-$AB \neq Y$ not-B and Not-$AB = Z$ not-A are equivalent, i.e. Not-$AB \neq$ not-AB not-B and Not-AB = Not-AB not-A are equivalent; for 'not-AB' put 'X' on one side. For Not-AB contains at least one of 'not-A' or 'not-B'. So if it does not contain one, it will contain the other; which, however, does not exclude that it contains both.

(17), (18), and (19) all represent variants of the principle of contraposition. Just like (18) follows from (17) by substituting '$\sim C$' for 'B', so conversely (17) follows from (18) by substituting '$\sim B$' for 'C'. Furthermore, (18) (and hence also (17)) is easily derivable from (20), because, e.g., the theorem $A \sim C \in\sim C$ (**Conj 3**) entails $\sim\sim C \in\sim (A \sim C)$ (by **Neg 2**) and therefore $C \in\sim (A \sim C)$, i.e. (18).[56] As Leibniz attempts to show in (19), the "normal" version of the principle of contraposition, i.e. ($A \in B \leftrightarrow\sim B \in\sim A$), or ($A = AB \leftrightarrow\sim B =\sim (AB)$), conversely follows from (17). Leibniz first derives ($A = AB \rightarrow\sim B =\sim (AB)$) from (17): $\sim B =\sim B \sim (AB)$, by substituting 'A' for 'AB' (which is permitted because of the premise $A = AB$) thus obtaining $\sim B =\sim B \sim A$. Next Leibniz tries to derive the converse implication in a similar way from (17), but here he encounters a difficulty. The premise $\sim B =\sim B \sim A$ in conjunction with (17), $\sim B =\sim B \sim (AB)$, immediately entails that $\sim B \sim A =\sim B \sim (AB)$; but from this equation one may not simply "subtract" '$\sim B$' to obtain $\sim A =\sim (AB)$ (from which the wanted conclusion $A = AB$ would easily follow by **Iden 5**). Leibniz himself was well aware of this problem and reminded himself that a "subtraction" is not *generally valid* but holds only in case of "uncommunicating" terms.

In the final (20), Leibniz once more comes very close to discovering the law **Disj 3**, $\sim (AB) \in\sim A \cup \sim B$. But – let it be repeated – since he doesn't have the

56 The idea of this proof was outlined in the *GI*; cf. §§ (76a) and (77).

operator of conceptual disjunction at hand, he "has to" paraphrase this by means of propositional disjunction as

Neg 21* $\sim(AB) \in \sim A \vee \sim(AB) \in \sim B$.

As was pointed out by Couturat in 1903, **Neg 21*** is invalid.[57] Actually Leibniz himself had noticed the invalidity of this principle already in 1686.[58] Hence his search for the "Fundamenta calculi logici", which he carried out in August 1690, was not crowned with success. The derivation of **Neg 2** from (17) remained incomplete; and even if he had found a complete and sound derivation[59], this wouldn't have been of great help since, as Leibniz noted above, the premise, (17), still "remains to be proved in our calculus".

The last quotation exhibits the greatest weakness of the brilliant logician – he was *too ambitious*! Leibniz wanted to *prove* the principle of contraposition while it should rather be considered (or, indeed, "*has to*" be considered)[60] as an *axiom*, i.e. a fundamental principle needed for proving other theorems but not itself in need of being proved. It is a certain irony of fate that, by 1690 at the latest, Leibniz had discovered all "Fundaments of the logical calculus" but still believed that he had not yet achieved this goal.

Even though, then, the essays of 1^{st} and 2^{nd} August are not masterpieces in the sense of finishing or rounding off Leibniz's search for the fundaments of a universal calculus of concept logic, they are apt to support Bertrand Russell's opinion on Leibniz:

> Apart from his eminence as a mathematician and as the inventor of the infinitesimal calculus, he was a pioneer in mathematical logic, of which he perceived the importance when no one else did so.

---

57 Cf. *Cout* 423, fn. 3. The invalidity becomes even more obvious if one transforms **Neg 21*** (by contraposition) into $A \in AB \vee B \in AB$.

58 Cf. *GI*, § 105: "If A contains not-(BC), it doesn't therefore follow that either A contains not-B or A contains not-C [...]".

59 E.g., Leibniz might have argued that since one »half« of the principle of contraposition, $A \in B \rightarrow \sim B \in \sim A$, has already been proved, the other "half", $\sim B \in \sim A \rightarrow A \in B$, is easily obtained by substituting '$\sim B$' for 'A' and '$\sim A$'; one thus obtains $\sim B \in \sim A \rightarrow \sim\sim A \in \sim\sim B$, hence by the trivial **Neg 1** $\sim B \in \sim A \rightarrow A \in B$.

60 Of course, whether or not a certain principle must be assumed as an axiom depends on the remaining axiomatic base of the calculus. If, e.g., Leibniz would have been willing to accept **Poss 4b** as an axiom, then **Neg 2** might have been proved as follows. If $A \in B$, then by **Poss 4b** $\neg P(A \sim B)$, hence $\neg P(\sim BA)$ or $\neg P(\sim B \sim\sim A)$, i.e. (again by **Poss 4b**) $\sim B \in \sim A$.

> He did work on mathematical logic which would have been enormously important if he had published it; he would, in that case, have been the founder of mathematical logic, which would have become known a century and a half sooner than it did in fact.[61]

And the present investigations hopefully also support the following claim of mine:

> Leibniz is the most important logician between Aristotle and Frege […] Leibniz's logical ideas were so far ahead of his time that even at the beginning of the 20^{th} century they remained, almost inevitably, misunderstood.[62]

5 Abbreviations of Leibniz's Works

Acad = German Academy of Science (ed.), G. W. Leibniz, *Sämtliche Schriften und Briefe*, Series VI, Philosophische Schriften, Darmstadt 1930, Berlin 1962 ff.

Cout = Louis Couturat (ed.), *Opuscules et fragments inédits de Leibniz*, Presses Universitaires de France 1903; reprint Hildesheim (Olms) 1961.

GI = *Generales Inquisitiones de Analysi Notionum et Veritatum*, first text-critical edition by Franz Schupp, Meiner 1982; cf. also Acad vol. 4, 739–788.

LLP = G. H. R. Parkinson (ed.), *Leibniz Logical Papers — A Selection*, Oxford Clarendon Press 1966.

61 Cf. Russell (1946, p. 618 and 613/614).

62 Cf. Lenzen (1983, pp. 418–419).

Winfried Löffler

Multiple Religious Belonging: A Logico-Philosophical Approach

1 Introduction: A typically Western problem? And whose problem?

In Austria and various other European states many official forms, e.g. for getting a certificate of registration or enrolling a child in a school, contain a box to fill in one's "Religious Confession". This box is usually rather small, it would hardly be possible to fill in more than one religion, and there is also no need for that: to most people the very idea of filling in more than one religion would sound absolutely queer. It goes without saying for almost anyone that, if at all, one can only belong to or be a member of exactly one religion.[1] Multiple religious belonging (i.e., roughly spoken, the "adherence" or "belonging" to more than one religion), hence, seems so obviously problematic that it is more or less a non-topic in the Western World. There are of course some remarkable personalities who participated in more than one religion and whose live and fate found high public attention. Many of them are Roman Catholics (Antony de Mello, Raimon Panikkar, Henri Le Saux, Hugo Enomiya Lasalle, Paul F. Knitter etc.), and a good part of this attention goes back to the problems which they had with the official Catholic church. This only underpins that multiple religious belonging is seen problematic in the West. According to the distinguished theologian and expert for Asian religions Peter C. Phan, however, things are very much different in Asia: multiple religious belonging [MRB] is rather the rule than the exception in certain regions of Asia[2], since especially local religious identities often draw from various religions and many people feel in a natural way connected to more than one tradition. (Whether this is always appropriately called "multiple religious belonging", or whether it is rather a form of "practical inclusivism" will be a matter of discussion in sections 3 and 7).

1 Explorative telephone calls in spring 2016 to Austrian authorities which are in charge of such issues confirmed the suspicion that the topic is nonexistent there. It was partly even hard to make the question understandable and to keep away the suspicion of a mere hoax call.

2 Phan (2003, p. 498)

Winfried Löffler: Innsbruck Universität

DOI: 10.1515/9783110529494-016

But is MRB a topic for the philosophy of religion at all? Other disciplines, or so one might be tended to think, could be more apt to address it, especially theology, sociology, psychology and history of religion. *Theological* questions concerning MRB might be those (and they are of a normative, not descriptive kind): As members of XY (Buddhists, Christians, …), what should we think about other religions and their members? E.g., what is the function of other religions (perhaps their function in a larger-scale divine plan), and is there something like salvation possible for their members? In the Christian realm, questions of that fashion are often labelled as a "theology of the religions". Some questions of the *sociology of religion* concerning MRB might revolve around what forms of MRB exist, how numerous they are in fact, whether they grow in number etc. These questions are descriptive in nature, similar to the questions which the *psychology of religion* might pose: How do people living with MRB feel and think? E.g., do they feel enriched or do they, perhaps for a certain time, also suffer under a sort of bad conscience? How can processes of adopting MRB be described and explained, be it in cases of an individually and newly adopted MRB, or in cases where MRB is the standard product of religious socialization? Likewise, the *history of religion* might raise descriptive questions about individual and collective MRB phenomena and their developments in past and present.

What, finally, might be relevant *philosophical* questions revolving around MRB? (Is there anything left to do at all for philosophers of religion, given the previously mentioned and pretty long list of issues discussed by scholars from other disciplines?) As philosophers are experts in questions of explication and the clarification of concepts, such questions might obviously be these:

- What concepts (e.g. what concepts of "religion", "belonging" etc.) are tacitly in the background?
- Especially: What is the conceptually appropriate way to address the topic? In the (in sum rather scarce) literature, the most frequent candidates are "multiple religious belonging", "multiple religious participation", and "multiple religious identity". But sometimes there is a worry being exposed whether the former concepts might be too static and too much focused on religions with clear members/participants – the Abrahamitic religions Judaism, Christianity, and Islam might be too much in the foreground here. "Multiple religious identity" is proposed as a more open notion in that respect.[3] The question of appropriate concepts of course connects to descriptive questions of religious studies here: it cannot be answered without a deeper look at the self-understanding of believers of different religions. (Throughout this paper

3 See, e.g., Bernhardt and Schmidt-Leukel (2008)

I will opt for "multiple religious belonging" as my working-concept. But this is for mere pragmatic reasons, and there is no theoretical claim connected with that option. All that will be said can also be reformulated using another concept.)

- There are some concepts in the semantic neighborhood of MRB which are not to be confused with MRB, especially "syncretism", "patchwork spirituality / religiosity", "inclusivism / inclusivist religiosity", "conversion". Nevertheless, a part of the literature seems to burdened with unclear conceptual borders here. Again, explications might serve as a means of clarity here.
- Lastly, philosophers of religion might ask the question whether MRB – in a certain understanding of the term – is a *rational position*.[4]

In this paper, I will first propose that many of the philosophical questions around MRB might ultimately go back to ambiguities in the notion of "religion" (section 2). This leads to a clearer analysis of the various questions arising around MRB and it is proposed that the greatest obstacles for MRB are to be expected around the theoretical conflicts between religions (section 3); the logical reasons for that are easy to identify (section 4). A promising conceptual tool for estimating the chances for rational MRB is the familiar distinction between exclusivism, inclusivism and pluralism which can be adapted for our tasks. As one might expect, an exclusivist standpoint would make rational MRB impossible and a pluralist one rather easily possible (section 5). On inclusivist backgrounds, however, the question for MRB suggests a couple of logical distinctions, and on the whole the prospects for rational MRB appear astonishingly poor (section 6).

2 The background of the problem: what is "religion"?

There is a wide consensus about the fact that there is no consensus in contemporary religious studies about an adequate definition of "religion". As a

4 In ch.2 of my *Einfuührung in die Religionsphilosophie* (Loeffler (2006)) I defend the claim that the question for the (ir-)rationality of religious beliefs should be regarded as the central task of philosophy of religion; questions for the nature of religion and the religious mind, the functions of religious speech and the structure of religious explanations are of course important, but they are just subsidiary questions to the question for (ir-)rationality.

consequence, many scholars today have more or less discarded this question (which was highly topical some time ago). The reason is basically that "religious" phenomena are too manifold to admit of a clear demarcation; sometimes even their adequate description is difficult. Nevertheless, there are still some defenders of "essentialist" and "functionalist" attempts to define religion. According to "essentialists", a "religion" is anything that displays a certain essential core (and "the assumption of certain world-transcendent realities like gods, spirits etc." is a popular candidate which is often proposed as this core). However, this has the somewhat awkward consequence that certain forms of Buddhism would not be religions. And the general question might be raised whether any proposal made here can ever be free of the perspective of certain religions which serve as paradigms. It might just be a prejudice to see this reference to transcendence as the core of religion; but such prejudices are not harmless since they might blindfold us for other important aspects of the religions. "Functionalists", on the other hand, define "religion" as anything that fulfils certain functions, be it individually (e.g. creating a sense of life and stabilizing the psyche) or socially (stabilizing certain aspects of social life). An obvious problem of such functionalist definitions is that they tend to be too wide; political ideas, sports, music, fashion, consuming goods, engaging in social and charity activities etc. may have similar functions, yet they are not religions. Furthermore, functionalist views fit best to traditional, religiously homogeneous societies: for such societies, it appears plausible that religions contribute to their stabilization. On the background of pluralist, partly secular societies, however, it is not obvious what is being stabilized by religions. What, e.g., is being stabilized by the various churches and religious associations in a religiously pluralist state like Germany or the United States? Some religions even take a clear critical stance towards the political and economic mainstream in many countries (think of the Catholic Social Doctrine, or – as a more radical example – religious terrorism); what is being stabilized by such (peaceful or violent) critical forms of religion? The possible reply that, e.g., the religious terrorist contributes to stabilize a fundamentally different conception of social life seems rather far-fetched and artificial.

Given these shortcomings of essentialist and functionalist definitions, it might be more appropriate to content oneself with a (broadly Wittgensteinian) exemplary definition of "religion": Although there is no general definition of "game" (since games are too manifold in their characters), the notion of a "game" can be introduced in sufficient clarity by reference to some doubtless examples of games and the various family-resemblances between these examples and other cases. Likewise, it should be sufficient to begin with some doubtless and well-understood examples of religions, list up some of their features (which they

might, however, display in different degrees) and take this list as a guideline to identify other religions via family-resemblances.

Religions as complex phenomena usually display the following features:[5]

- They are realized in a social group with more or less clear borders/memberships and inner social structures;
- within this group there can be "religious experts" or "functionaries" with special roles, abilities, or competences (think of priests, munks and nuns, media, prophets, templewardens and the like);
- there can be festivities, rituals, and other more or less formally fixed procedures;
- many religions know significant ("holy") places, times, and objects, there can be "taboos" like untouchable things or persons, places forbidden to enter, forbidden kinds of food or drinks, etc.
- religions contain some (more or less elaborate) theory-like core of claims ("God created the world", "God exists and is a Trinity", "all conceptualizations of ultimate reality are illusions", …)
- religions offer a (more or less elaborate) world-picture with descriptive and evaluative aspects, e.g. they make proposals about what parts of "reality" do exist, where we humans come from, where we will ultimately go, what goals are "really" important in life etc.
- religions comprise a more or less detailed special code of behavior, which may contain individual moral commandments as well as commandments for social behavior;
- many religions expect (and partly cultivate) extraordinary states of the body and/or the mind or extraordinary forms of communication (like meditation, prayer, ecstatic states etc.)

It should be stressed that different religions may put a very different emphasis on these features (even if probably any religion displays at least something from each point this list). E.g. there are highly ritualized religions as well as others who focus more on the moral behavior of their members, or others who are in the first place an offer of a world picture. It may also be that religions change their emphases in the course of their history (e.g., Catholicism before the 1960s put much more emphasis on the exactness of detailed rituals than it does today; conversely, Protestantism (at least in Central Europe) today has a tendency to appreciate rituals somewhat higher than in the past). We shall see that these variances between different religions also influence the possibilities of MRB.

5 The list is not exhaustive. See also Loeffler (2006) (footnote 4), 9–18.

3 Where MRB problems appear

3.1 According to the Indonesian theologian Albertus Bagus Laksana[6], discussions about MRB revolve around one or more of the following three dimensions:

(1) around the practice of meditation;
(2) around questions of intellectual / theological assent;
(3) around questions of institutional affiliation.

If we take these three dimensions as the starting-point of a more general reflection, we see that all the above-listed features of religion, can, in a loose sense, also be grouped along these three dimensions; they fall into:

(1*) practical features: rituals, festivities, holy places, times, objects, taboos etc., extraordinary of body states / mind states / communication forms (prayers, meditations, ecstatic states, …);
(2*) theoretical features: theory-like core; World-picture, descriptive and evaluative;
(3*) social features: social groups with memberships, structures, "experts", functionaries; religious code of behavior.

Laksana's list appears somewhat overly specified (e.g., MRB might not only become a topic concerning the practice of meditation, but also concerning rituals, visiting holy places or observing religious holidays); but it is illuminating since we could take it as standing *pars pro toto* for the three groups of features of religion. Hence, we shall in a somewhat more general sense say that MRB problems may appear on the level of the practical, the theoretical and the sociological aspects of religions.

3.2 As a general hypothesis, I propose that *it is the level of the social features (3*) and (especially) of the theoretical features (2*) which tends to create problems or obstacles for MRB*. In somewhat more detail, one could presume that:
- The more a religion emphasizes "practical" aspects, the easier is MRB also to other religions;
- the more a religion emphasizes its "theoretical" aspects, the more difficult is MRB;

6 Laksana (2014)

- the clearer-cut the social aspects are (e.g., memberships and internal structures), the more obstacles for MRB are to be expected;
- the more a practical feature is separable from theoretical ones, the easier is multiple religious participation in it (examples might be inter-religious activities to help the poor, or inter-religious New Year celebrations);
- MRB can become especially difficult where religions have clear-cut membership conditions and/or initiation rituals with an "exclusive" implication (e.g., according to the standard interpretations of the Christian baptism or the Islamic recitation of the creed (Shahada) for the first time, these rituals imply a renunciation of all rival religious affiliations.
- One might add, for the sake of precision, that full-blown MRB should be distinguished from a sort of "practical inclusivism" (i.e. the constellation where persons would declare some religion to be their primary one, but include various (especially) practical aspects, e.g. festivities, rituals and even some beliefs from other religions into their own special shaping of religion. The difference to MRB is that such people would presumably not feel as members of more than one religion.

3.3 It is not always clear from the outset whether a phenomenon or a social practice should be labelled as "(non-)religious", and moreover these labelings are not always uniform across different religions. For example, venerating the family ancestors would be seen by many Chinese and Vietnamese people not so much as a religious practice, but as a normal aspect of appropriate and decent social behavior. Consequently, this practice is observed, e.g., also by many atheist party members, ancestors' altars are found also in Christian Vietnamese families, etc. Nevertheless, an important aspect of the Chinese rites controversy during the 17th and 18th centuries between the Vatican and the Jesuit missionaries revolved around these practices. The example might be insightful for the relation between theoretical and non-theoretical features of religions; we shall return to that relation in the next section.

4 The logical reasons of the problem

We have seen so far that MRB tends to cause problems when features on the social level and especially on the theoretical level are involved, i.e. when person somehow have to decide between social and/or theoretical options. But why is that so? There is a simple logical reason for that: It is only theoretical claims that can be contrary or contradictory to each other directly. It is usually only

for theoretical claims that it makes sense to construe a decision situation of the structure "P or non- P?"; where social or especially practical aspects are at stake, apparently similar decision-situations might go back to a connection with theoretical interpretations. This claim will be elucidated in what follows.

With few exceptions, practical actions (seen as such) cannot be contrary or contradictory. E.g., lighting candles in a temple and in a church is as such not contra(dicto)ry; celebrating Chinese and Western festivities is as such not contra(dicto)ry; trying to learn from holy persons of different religious traditions is as such not contra(dicto)ry. (There are, of course, certain metaphysical incompatibilities between actions: e.g., one cannot go to temple and church at the same time, etc. But for the present topic of MRB such incompatibilities are irrelevant.) Likewise, social features or actions (seen as such) are not contra(dicto)ry: For example, being a member of two or more religions, being a Christian monk and a Zen Master at the same time is as such not incompatible. The problems arise when practical actions and social actions or features are connected with specific theoretical claims from one or both involved religions. Within a certain theoretical framework, practical actions and social features/actions can become appear contra(dicto)ry. E.g., a theoretical claim like "lighting candles in a holy place means tacitly subscribing to all the doctrines which are taught there" might turn a rather innocent practice into a veritable interreligious problem. (To name an example, the famous 17th/ 18th century debates about the activities of Matteo Ricci S.J. and other missionaries to China were also burdened by claims like "allowing some Chinese rituals to Chinese Christians means denying God's uniqueness"). Apparently, some initiation rituals which are connected with declaring certain theoretical beliefs (or rejecting some other beliefs) put up special obstacles for MRB, when these declarations / rejections create logical incompatibilities with other religions. The connection with theoretical beliefs might be patent and obvious (as in the baptism of an adult, where the catechumen is explicitly asked for certain beliefs), or it might result from a speech-act-theoretical analysis of the ritual situation.

5 MRB and theoretical conflicts between religions: three options

5.1 So far we have seen that MRB is most easily possible on the practical side of religions; and that difficulties can arise where practical aspects are connected with theoretical claims, and where clear membership borderlines are connected with certain requirements of assent to theoretical contents. Seen from that angle,

we might presume that in general MRB is easier for many Eastern religions (with their focus on practical aspects, and few clear initiation rituals), and that a lot of de-facto-MRB works smoothly without being a problem or without even being noticed. In that sense, Laksana's[7] claim that "local cultural spaces" (with their special mix of traditions) are more important for religious identities in Asia than the adherence or the feeling of belonging to certain religions seems plausible. Conversely, it appears that MRB tends to be difficult in religions with a strong emphasis on theoretical aspects and clear membership and initiation requirements, as in many Western religions. It his hence no surprise that many of the prominent and notorious personalities in the MRB discourses (see section 1 above) were Catholics who faced many obstacles on their personal way.

5.2 Still, the foregoing considerations do not fully exclude MRB also in cases where "theorycentered" religions are involved, and there are also some credible examples for it. But in what sense is it logically possible and to what extent is it rational?

The answer to these questions depends on the stance which a person takes in the question of the logical relations between different religions. A plausible conceptual approach to tackle this question is the usual and well-entrenched distinction between exclusivism, inclusivism and pluralism, but it seems helpful to begin with a qualification here. The distinction can be understood in two different ways which are not always clearly distinguished: firstly in an epistemological (or, more specifically) truth-related way (i.e. as revolving around the truth claims of one (or more) religion(s)), and secondly in a salvation-related way, as a distinction of positions about the prospects of an ultimate salvation (or other forms of ultimate fulfillment) for the adherents of one (or more) religion(s). Salvationrelated exclusivists see a possibility of salvation only for adherents of one religion, inclusivists see it also for adherents of a range of other religions (perhaps with more difficulties, and provided they have sufficient similarities to the own religion), and pluralists see it for the adherents of all religions. The distinction is probably more often taken as a salvation-related one, but in the present context it will be understood in the epistemological, truth-related way.[8] I.e. will ask for the rationality of MRB in cases where religions with a

7 See footnote 6 above.

8 Various combinations of the two readings are thinkable. E.g., one can be an exclusivist concerning truth and yet have an inclusivist or even pluralist conception concerning salvation: One might, e.g., think that only one religion contains true religious propositions, but that the doctrine of this religion admits of a possible salvation of members of some or perhaps even all other religions, their entirely false doctrines notwithstanding. This is in itself not inconsistent;

strong theoretical side are involved – and this before the different backdrops of exclusivist, inclusivist and pluralist standpoints concerning the truth-claims of different religions.

5.3 Truth-related *exclusivists* hold that only one religion contains true religious propositions, and that hence all the others contain can only false religious propositions.[9] At closer look, such a position is in fact a very demanding one, rather a matter for strong religious fundamentalists and probably hardly ever defended seriously. E.g., it would require that even similar-sounding propositions of other religions are *entirely false*. If two religions *A* and *B* hold doctrines, e.g., about some form of divine creator of the world, then an exclusivist interpreter of *A* would have to consider the *B*-beliefs about the creator as entirely false – i.e. not even as partially true, vaguely true, analogically true, verisimilar (or as containing a grain of truth in some other form).[10] It is easy to see that MRB (if "belonging" implies the acceptance of at least one religious proposition of the respective religion[11]) would appear as epistemologically irrational.

5.4 Truth-related *inclusivists* hold that only one religion grasps the full truth, but that other religions may well contain many partial truths and valuable insights. Inclusivism seems to be the position of Christian and Islamic mainstream theologies, and also e.g. many Buddhists share it.[12] It might seem that from an

---

the appearance of inconsistence emerges only when salvation is believed to be connected with believing some or all of the true religious sentences.

9 The adjacent problem of the proper demarcation between religious and non-religious propositions is only mentioned, but not further investigated here. Exclusivists would probably not incline to hold that members of other religions are wrong in *all* of their claims, be they religious or non-religious.

10 Presumably, most purported "exclusivists" would at closer investigation rather turn out as inclusivists who attribute only very few (and/or perhaps very general and/or unimportant) truths to other religions.

11 The considerations in this and the following chapter are in many points inspired by the first part of Bernard Bolzano's *Lehrbuch der Religionswissenschaft*, here especially §§20–22. Bolzano's significance as a forefather of analytic philosophy of religion is still widely underrated.

12 A classical Catholic inclusivist statement is the Declaration "Nostra Aetate" (1965) of the Second Vatican Council, especially ch. 2: "From ancient times down to the present, there is found among various peoples a certain perception of that hidden power which hovers over the course of things and over the events of human history; at times some indeed have come to the recognition of a Supreme Being, or even of a Father. This perception and recognition penetrates their lives with a profound religious sense. Religions, however, that are bound up with an advanced culture have struggled to answer the same questions by means of more refined concepts and a more developed language. Thus in Hinduism, men contemplate the divine mystery and express it through an inexhaustible abundance of myths and through searching philosophical inquiry.

inclusivist standpoint the prospects for rational forms of MRB are better. However, things are overall not as simple as they might initially appear; we shall hence postpone the discussion of the possibilities of MRB under inclusivist assumptions to a section of its own (sect. 6).

5.5 Truth-related *pluralists* hold that the doctrinal corpora of all (or at least: of all the major or more important) religions are equally true. Such a claim would of course require some conceptual and/or logical devices to neutralize the apparent contradictions between the religions. A usual approach to do this is to declare the various religions as different versions of telling the same existentially important truth about the central object of religion (a famous and well-known way to put this is John Hick's thesis that the (transcategorical) "Real" manifests itself in various forms of appearance[13]). Provided that these logical problems can really be overcome, it seems obvious that MRB would be possible and not particularly difficult from a pluralist standpoint, since the theoretical clashes between religions vanish.[14] But a requirement for MRB even under pluralist assumptions is of course that the social aspects of the involved religions (including the "self-understanding beliefs", see 6.1 below) are not, e.g., perceived as leading to mutual exclusions between religions. (Another question is of course whether the pluralist account as such is plausible; we shall come back to this question in section 7.2 below.)

They seek freedom from the anguish of our human condition either through ascetical practices or profound meditation or a flight to God with love and trust. Again, Buddhism, in its various forms, realizes the radical insufficiency of this changeable world; it teaches a way by which men, in a devout and confident spirit, may be able either to acquire the state of perfect liberation, or attain, by their own efforts or through higher help, supreme illumination. Likewise, other religions found everywhere try to counter the restlessness of the human heart, each in its own manner, by proposing "ways," comprising teachings, rules of life, and sacred rites. *The Catholic Church rejects nothing that is true and holy in these religions.* She regards with sincere reverence those ways of conduct and of life, those precepts and teachings which, though differing in many aspects from the ones she holds and sets forth, nonetheless often reflect a ray of that Truth which enlightens all men. Indeed, she proclaims, and ever must proclaim Christ "the way, the truth, and the life" (John 14:6), in whom men may find the fullness of religious life, in whom God has reconciled all things to Himself." (My italics.)

13 Hick (1992)

14 A more radical variant of truth-related pluralism would of course suspend all realist claims connected with religious language. Religious claims, under such a view, would not refer to some objective transcendent reality which is described from different perspectives, but they would just be internally true if uttered appropriately within the respective religious language-systems which have no interesting relations between them. MRB would of course easily be possible here. However, I fear that such an understanding of religion is at odds with the semantic intuitions of most religious believers; hence it will not be further investigated here.

6 MRB under inclusivist assumptions: A surprising nest of logical problems

Let us now analyze the prospects of rational forms of MRB under inclusivist assumptions. Although inclusivism will *in itself* be considered the preferable option of the three (see 7.2 below), it will turn out that the possibilities for (rational) MRB under inclusivist assumptions are less promising than one might perhaps expect. More precisely, we shall see that some forms of MRB seem theoretically easily conceivable, but they appear practically unrealistic (6.2), whereas the more realistic and relevant forms of MRB will turn out surprisingly difficult to specify (6.3): At closer inspection, although there is a huge spectrum of possible logical relations between religious belief systems, not too many of them endorse genuine and rational MRB. From a logical standpoint – so one might summarize – things are not as simple as they might initially seem, given the overall open and MRB-friendly appeal of inclusivism.

6.1 *Three preliminary remarks on religious belief-systems and adherence*: The following considerations are based on a couple of idealizing assumptions and provisos.[15] A first one is that the actual sets of beliefs of the individual adherents correspond to the so-to-say "doctrinal standards" of their respective religions (or at least the average/mainstream sets of beliefs; we come to that point in a minute). I.e., we abstract from the fact that in practice people sometimes arrange their highly individual and non-standard sets of the beliefs they accept (or which they know of and/or which they find relevant). Under this assumption we shall ask for the prospects of logically consistent and rationally acceptable forms of MRB given that the involved persons have an inclusivist standpoint concerning the truths of religious beliefs.

Another qualification is in place concerning the exact content of these "doctrinal standards". One might tend to believe that the doctrinal standards of religions are simply to be equated with the *entireties* of their respective doctrines, and adherence would be simply defined by assent to these entire belief-sets. But it was already Bernard Bolzano in the early 19th century who rightly observed that there is a certain variety here and that the religion of a group is best defined as the

15 These assumptions are not especially connected with inclusivism. In principle they would also be relevant for marginally more precise accounts of pluralism and exclusivism and the possibilities of MRB there (hence they could also have been introduced in section 5 above). However, since their impact and relevance is better seen in the context of inclusivism, they are being introduced here.

set of propositions that is believed by *most* of the members of a group.[16] Bolzano said this on the background of a comparatively clear-cut religion like Catholicism (with its Catechisms etc.); a fortiori it should be kept in mind for religions whose stock of doctrines is fuzzier in character. It is hence not belief in the *entirety* of doctrines which should be seen as defining the adherence to a religion, but rather belief in a mainstream or average set of them. Where set-theoretical models will be used below, the outer circles will hence represent such average sets of beliefs.

Moreover, the related phenomenon of unequal weights of propositions should be taken into account: not every doctrinal proposition of a religion has the same importance and relevance for the adherence to a religion. This can be seen, among others, by the role of creeds: creeds summarize the most important doctrinal propositions assent to which is often regarded as necessary for adherence.[17] In the models below, there will hence be a (simplified) distinction between core propositions (represented by the inner circles) and the average set of propositions (outer circles), and the former are always assumed to be subsets of the latter.[18]

6.2 *A first (but rather unrealistic) scenario of inclusivist MRB – total inclusion*: A general distinction for inclusivist scenarios seems to be the following: the doctrines of a purported "including" religion can contain the doctrines of an "included" one *completely*, or there might only be a *partial* overlapping between the two[19] involved religions. Let us begin with the theoretically conceivable case that one religious belief-system *A* is *completely* "included" in another system *B*, i.e. that the doctrinal propositions of *A* constitute a subset of *B*'s doctrinal

16 Bolzano, loc.cit., §22 n.4.

17 Catholic theology explicitly talks about a "hierarchy of truths" (see, e.g., *Catechism of the Catholic Church* (1992), § 90), but similar views exist also in other religions. In Islam, the affirmative recitation of the shahada, a sort of minimal creed ("There is no god but God. Muhammad is the messenger of God.") is even constitutive for membership. – There is of course a mass of possible weighting problems lurking here which I can only mention: Should a person who fails to believe in one member of the core-set, but believes in the whole average set be considered a non-adherent? If no, should she be considered more/less an adherent than someone who believes only in the core set, but not more? Can lacks in beliefs in the core be outbalanced by numerous beliefs outside the core? Does adherence admit of degrees at all?

18 This is another idealizing assumption: What almost all members of a religion believe can be more, less or even something different from the core propositions of that religion. I do not address the questions of how and by whom the core propositions are singled out. Religions need not necessarily have a clear-cut procedure or social mechanism to do that.

19 In order to keep things simple, we consider only the case of two involved religions.

propositions, and that hence all adherents of *B* should eo ipso also be adherents of *A*.[20]

Given our assumptions above, this theoretical case should now be somewhat specified: if what most adherents of *A* believe is a subset of what *most* adherents of *B* believe, and if also the core propositions of *A* are a subset of the core propositions of *B*, and if most adherents believe in their respective core propositions, then most adherents of *B* should also be adherents of *A*. And this would hence (in a sense) constitute a form of MRB. However, it appears doubtful whether relevant examples for such a form really exist. Of course, quite clear-cut sub-groups of certain religions do indeed exist, sometimes they define themselves by additional doctrinal propositions which go beyond the mainstream doctrines. (E.g., there are Catholic Christians whose religious life is centered around the messages from certain controversial visionaries and/or places of pilgrimage, i.e. they have a richer set of core of propositions.) In principle, regarding the structures of belief-sets alone and leaving aside the aspect of self-understanding, such a scenario would constitute a form of MRB would appear as rational. But the adherents to such sub-groups would presumably not understand themselves as adherents to a religion *B* of its own in addition to their religion *A*, but rather as adherents to *A* with a special addendum or a special focus on certain doctrines.

The foregoing considerations might remind us that there is a special ingredient in the belief-sets of most adherents of religions which might in general considerably reduce the prospects for genuine MRB: the belief that describes the self-understanding of a person as an adherent of this or the other religion. Let us call it, for short, the self-understanding belief. It is not necessarily a matter of a private, subjective commitment, it might also well be a part of a collective understanding of identity ("we *A*-adherents in distinction from the adherents of *B*, *C*, ..."). And one might presume that this self-understanding belief ranks among the more important beliefs in the belief-system of a religious person. We shall come back to these self-understanding beliefs later on.

6.3 *Inclusivist MRB scenarios with partial overlapping*: The case that the sets of doctrinal propositions of the involved religions do not entirely, but only partially overlap is the practically more relevant one, and it appears to open more logical space for MRB, at least prima facie. Given our foregoing considerations, we might construct a variety of scenarios here. Again, we consider the case of

20 Those who find this claim prima facie counterintuitive are reminded that someone who believes in to a bigger set of doctrines *a fortiori* believes in a smaller one, but not conversely. Hence the number of adherents shrinks with the number of accepted propositions.

only two involved religions *A* and *B*, and we stepwise introduce qualifications as before.

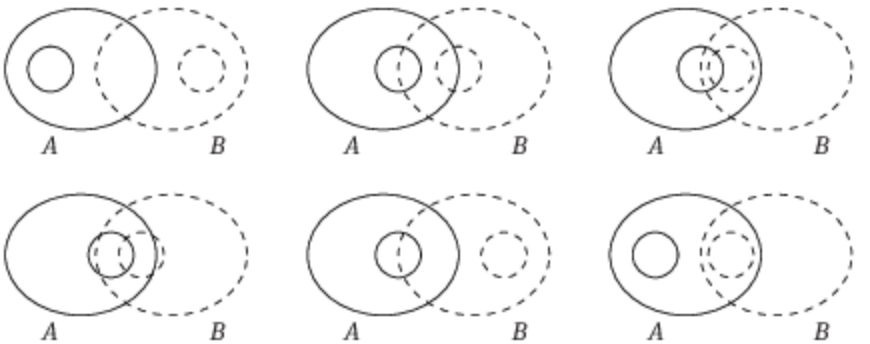
(i) The overlapping might comprise the *whole* sets of average beliefs (i.e. the sets of beliefs which are held by most members of the involved religions *A* and *B*), or this might not be the case. In the first case, it might seem hard to explain why we speak of two separate religions at all, since the nonoverlapping areas of the belief-sets will only contain non-standard additional beliefs of individual believers or groups of them. Nevertheless, such a case is possible (e.g., the de-facto overlapping between the entire "average belief-sets" of A and B might have gone hitherto unnoticed); hence, if the self-understanding beliefs ("we are *A*-adherents" and "we are *B*-adherents") are not seen as mutually exclusive, then MRB to *A* and *B* might be possible and rational. In the second case, however, i.e. when the overlapping does not comprise the entire "average belief-sets" of the involved religions, a purported MRB status would be – although psychologically possible – not entirely rational, since it relies on (partial) self-deception: if someone honestly believes to be an adherent of *A* and *B*, but fails to share the average belief-sets of (one of) these religions, and if "adherence" goes along with sharing standard beliefs, then she is not even fully rational in her beliefs about her respective adherence(s), and a fortiori in her beliefs about her MRB status.

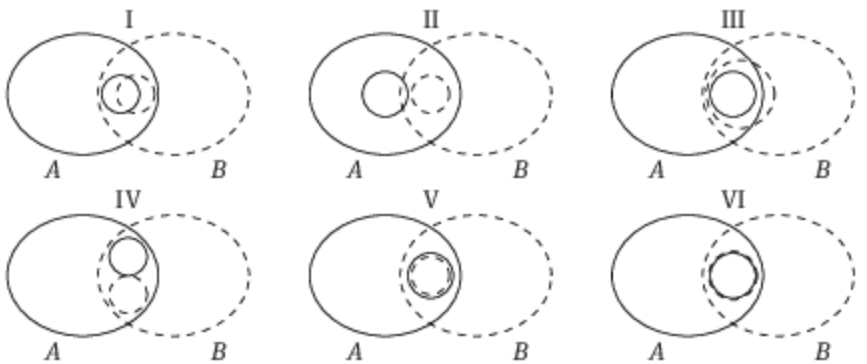
(ii) The aforementioned distinction between core propositions and other propositions of a religion complicates things also here. It is of course an empirical matter how the average belief-sets of religions de facto relates to the respective sets of their core beliefs: it is by no means excluded that the average belief-sets of the adherents omit certain core beliefs (as they are singled out, e.g. in creeds, catechisms, by renowned theologians, etc.) or even contradict them.[21] But let us, for the sake of simplicity, assume that the set of core beliefs is always a (proper or improper) sub-set of the average beliefs of a religion, i.e. that the adherents of a religion by and large believe in all the core propositions of their religion. (Maybe they believe also in more, but not in less and not in something different or something incompatible with the core beliefs.) The following diagram shows that even under such simplifying assumptions a lot of scenarios[22] are conceivable, of

21 Such differences do not seem far-fetched; see – for the Christian realm – the familiar laments about the hiatus between academic theology and the practice of faith and about the decline of religious knowledge.

22 In the following diagrams, I assume that the set of the core propositions of a religion are always a *proper* subset of the set of all its propositions, i.e. that every religion contains also some

which a non-complete[23] selection will be graphically modelled here (outer circles represent the average belief sets of the two involved religions *A* and *B*, inner circles represent their respective core beliefs; in the last scenario down right, the core beliefs of *A* and *B* are supposed to be identical):[24]





non-core propositions. By dropping this (probably realistic) assumption for one or both involved religion(s), some more scenarios are easily conceivable.

23 Throughout the models, I assume that the sets of average beliefs do always only partially overlap and neither do they coincide nor does one include the other. Dropping this (realistic) assumption would yield more possible scenarios. Moreover, many models (e.g. II and III in the box) would allow "mirrored" versions if A and B are exchanged.

24 Another – in fact bold – assumption behind the following diagrams is the inner logical consistency of the belief-sets of *A* and of *B* respectively.

If we reduce our demands and take it as sufficient for the adherence to a religion that all of its core beliefs are accepted,[25] then it is obvious that only in the six last (boxed) scenarios there can be forms of rational MRB; some of them by the adherents of both religions *A* and *B*, some of them only by the adherents of one of them.[26]

In sum, despite the open and inviting appearance of inclusivism which might raise the expectations also for MRB, our closer analysis suggests that rational MRB is surprisingly difficult to realize. At closer look, namely, the six remaining scenarios turn out quite demanding in practice (and hence probably not too realistic): Scenarios III and V assume that the core of one religion comprises also the full core of another (which raises the aforementioned problem whether the two religions will understand themselves as distinct at all, and not modifications of one).

Scenarios I and IV are less demanding in that point, but they assume that both involved religions do not see any falsities (or lacking points) in their cores: both cores are mutually accepted by both religions (see footnote 24 above). In practice, this is not too realistic, since different religions define and legitimate themselves usually by some differences in the cores.

Scenario II assumes this only for one religion: In our diagram, religion *A* would regard the entire core of religion *B* as a part of its non-core beliefs, but not (entirely) vice versa. This might perhaps appear as the less demanding and most realistic scenario, but it is affected by a coherence problem: The *A*-adherent – if she should really feel as an adherent of *B* as well – should probably regard the corebeliefs of B as core-beliefs. This does of course not raise a consistency problem (since we assumed in footnote 24 that everything that *A*-adherents believe is consistent), but a correlation or coherence problem: The A-adherent in scenario II would in fact have two sets of core-beliefs in her belief set without a guideline about how they correlate, how they hang together and how the internal

25 We need that assumption to make MRB possible at all: In our models, MRB is only possible where someone subscribes to the core beliefs (and perhaps also to the average beliefs) of one religion *and at least to the core beliefs of another*. Requiring more, i.e. requiring that she subscribes to both average sets, would either require believing in something she actually does not believe, or be a fallback into the model of total inclusion (6.2). – The assumption is one the one hand not unrealistic and on the other hand not a contradiction: Since the average set is only what most adherents believe, someone who only accepts the core beliefs does not thereby turn into a non-adherent.

26 This is an interesting result in itself: The rationality of MRB is non-symmetric in character: If it is rational for an adherent of *A* to be also an adherent of *B*, this does not imply that it is also rational for an adherent of *B* to be also an adherent of *A*. However, it can be rational.

weighting between them should be managed (core-beliefs, as we saw above, are important and have the function to structure and to guide a religious world-view; a multiplicity of incoherent sets of corebeliefs is hence at least pragmatically difficult to live with).

Scenario VI, finally, presupposes that two religions have completely identical cores. This invites, *a fortiori*, a similar objection to the one raised above against scenarios I and IV.

6.4 *Where inclusivist MRB tends toward pluralism*: Nevertheless, especially scenario VI (and perhaps, to a lesser extent also I, III and V) might have a certain initial attractivity for friends of the possibility of MRB. What different religions claim in their cores, or so the suggestion might go, is ultimately more or less identical. Hence, an MRB which also comprises the theoretical aspects should not be such a big problem. This position is of course viable, but it is probably at odds with the usual self-understanding of religions and their belief-systems, and it is hence not very realistic if taken as a still *inclusivist* position. In order to assimilate, e.g. the doctrinal core-beliefs of Christianity, Judaism, Islam and Hinduism so far that they can be regarded as identical, substantial modifications must be made (in the first place, probably, omissions and/or generalizations of core doctrines). The result of such an assimilation process, and an MRB based on it, would resemble a *pluralist* re-interpretation of these religions more than an inclusivist position.

7 Taking stock

7.1 In a nutshell, the insights of the foregoing chapters might be summarized as follows: MRB which does justice to the theoretical aspects of religion is impossible in an exclusivist framework, easily possible under a pluralist one and more difficult than one might expect under an inclusivist one (the latter two always with the proviso that the social aspects of the religions do not erect too many practical obstacles).

7.2 Hence, a complete answer to the question whether rational MRB is possible will have to include a commentary on the preliminary question whether the exclusivist, pluralist or inclusivist frameworks are plausible at all. It would transcend the scope of this paper to take a detailed stance in this decades-long debate, hence the following (unduly short) remarks do not intend more than to adumbrate the author's position. Exclusivism (as defined above in the demanding sense that only one religion can contain true sentences at all) seems doubtlessly false, given the obvious big similarities between many religions (many purported "exclusivists" will probably rely on a self-misunderstanding and in fact be

restrictive and non-generous inclusivists which grant other religions only few correct insights).

Pluralism seems false as well, however in a much less obvious way. From the many objections put forward in the literature,[27] I mention only two very striking ones: (a) Pluralism has to choose between the Scylla of substantial re-interpretation of religious doctrines (to make them mutually consistent in some restricted interpretation) and the Charybdis of applying a hardly intelligible and non-standard notion of the "truth" of religious claims. The "truth" of obviously inconsistent religious claims under a pluralist interpretation seems to be rather something like practical or spiritual utility for personal growth; in such a sense of "truth", linguistically inconsistent claims might indeed be true. (b) Some religious doctrines make explicit epistemological claims which are hard to reconcile with religious claims, but also hard to re-interpret into a "less troublesome" form. E.g. it is held in certain Buddhist traditions as a rather high-rank claim that "all conceptualizations of ultimate reality are illusions". There is no easy way at the horizon how to reshape this claim in a less literal way. But if this claim is a member in the set of claims which the pluralist has to take seriously, then not all claims can be true. Inclusivism seems by far the most viable and theoretically satisfactory position. It does justice to the similarities and differences of the religions in a plausible way. Much of what the pluralist claims in support of his position can also be seen as a support of inclusivism. E.g., there are strikingly similar spiritual values between many religions: the respect to and safeguarding of creation, the striving for humanity, wisdom and self-cultivation, the ideal of universal brotherhood, the preference for social order and the striving for a "sanctification" of personal and social life, and some more. This phenomenon can be seen as evidence not only for a pluralist, but also for an inclusivist standpoint. Similar spiritual values would be hardly explainable if there were not strong theoretical similarities between the religions. That the adherents of the individual religions naturally see their religions as the ones grasping the full truth (or having at least the optimal access to truth), need not be noxious. Especially, it can be sufficient for a peaceful cohabitation in religious tolerance and cooperation, if the "solution to the puzzle about the right religion" is left to God or the Ultimate Reality, and human beings do not arrogate the position as a violent executor of purported religious truths.

7.3 In sum then, the foregoing considerations suggest a rather limited space for rational and genuine MRB: A truth-related reading of the debate between exclusivism, inclusivism and pluralism suggests that pluralism (which would

27 For a similar position see, e.g., the clear exposition in Weidemann (2007), chapter 2.4.

provide the most promising background for rational MRB) is a much less plausible position than inclusivism. (Exclusivism can stay out of consideration anyway.) But under inclusivist assumptions, the possibilities for rational and genuine MRB turned out to be surprisingly marginal (or more precisely, they would hold for rather unrealistic accounts of religion only). Moreover, it should not be forgotten that even these rather meager possibilities rested on a number idealizing assumptions about religions which were invested to establish the rather simplified models used above. Presumably, things are much more complicated for "real life" forms of religion (which would further reduce the logical space for rational genuine MRB). All this suggests that many forms of purported "MRB" should at closer scrutiny perhaps better be described as mere forms of practical inclusivism or and/or of de-facto social multi-adherences without a full theoretical account behind them. This should of course not be read as an invitation to disrespect or devaluate such forms; even where their theoretical description and/or self-understanding might be deficient, they might still be valuable forms of enrichment from other religious traditions and important bridges between them.[28]

28 I am indebted to Christian Tapp for various constructive comments on an earlier draft of this paper.

Paolo Mancosu

Definitions by Abstraction in the Peano School

Abstract: In a previous publication Mancosu (2015a) I have discussed Frege's approach to definitions by abstraction by paying special attention to its Grassmannian roots and the teaching of elementary geometry in nineteenth century Germany (especially to debates on the concept of "direction"). Moreover, I emphasized the originality of Frege in taking the extension of a concept as the value of the abstraction operator thereby identifying, for instance, the direction of a line a with the extension of the concept "x is parallel to a". In this paper, I will take for granted the results of that article and I will only flag my use of them by referring to the article at the appropriate junctions. My attention in this paper will focus on the logico-foundational analysis of definitions by abstraction in the Peano school and in Russell. We will see that many of the philosophical debates at the time foreshadowed contemporary debates in the neologicist literature.

I dedicate this paper to Sergio Galvan. My intellectual debt to him is enormous. He taught me mathematical logic starting in my first year in college and he was also my undergraduate Laurea thesis advisor (on the topic of nonstandard models of arithmetic and Paris-Harrington incompleteness). Throughout the years our relationship has developed from one of professor-student into one of friendship and scholarly collaboration. My admiration for his originality, intellectual rigor, and depth has only increased with the years. I believe it is appropriate that I contribute a paper on the Peano school for this volume. Sergio still delights in recounting that already in my first year as an undergraduate I approached him suggesting a thesis topic on Peano's work in analysis and logic. He wisely steered me to the topic of non standard models of Peano Arithmetic, thereby giving me the impression that the connection to Peano was preserved even with the new topic! I guess my interest in the historical aspects of Peano's work was a lasting one since, after thirty-five years, it is still alive. This paper is part of a larger project on the history and philosophy of definitions by abstraction. I would like to thank Clara Silvia Roero, Paola Cantù, Jamie Tappenden, Marco Panza, Paolo Freguglia, and Ivor Grattan-Guinness for their help connected to Peano and his school, Russell, and Couturat.

Paolo Mancosu: University of California at Berkeley

DOI: 10.1515/9783110529494-017

1 Peano and his school[1]

In Mancosu (2015a), I pointed out that in 1887 Helmholtz referred to Hermann Grassmann in describing the process that leads from an equivalence relation to an identity of abstracta. His main examples came from physics (masses, weights, temperatures etc.). Of course, Frege in 1884 had also, but without using the terminology of abstraction, discussed the same type of concept formation (through definitions that we now call 'by abstraction'). In the same article I also pointed out that Frege's §64 in the *Grundlagen* is influenced by Grassmann and especially the essay *Geometrische Analyse* of 1847.

(Peano, 1888) contains the first description by Peano of definition 'by abstraction'. The terminology is not there yet (one has to wait to 1894 for the first full explicit use of 'definizione per astrazione' in a review by Vailati) but all the elements are in place. It is important to point out that the title of Peano's work is *Calcolo geometrico secondo l'Ausdehnungslehre di H. Grassmann*, which unequivocally shows Grassmann's influential role in shaping reflection on abstraction in the nineteenth century. In section 1 of this work, Peano defines equality between two entities of a certain system, written $a = b$, to mean a relation between elements of the system that satisfies symmetry and transitivity. Interestingly, he only states the properties but does not name them.[2] In section 80, Peano gives examples of relations that satisfy (1) neither symmetry nor transitivity; (2) symmetry but not transitivity; (3) transitivity but not symmetry; (4) both symmetry and transitivity. The latter are the important relations for definitions by abstraction. I recall that when a relation R satisfies that for every a there is a b such that aRb, reflexivity follows from symmetry and transitivity. His examples of relations that satisfy both symmetry and transitivity are:

1. the number a is equal to the number b;
2. the number a is congruent to a number b with respect to a fixed module;[3]

1 On Peano and his school see Borga et al. (1985) and Roero (2010). On Peano and abstraction with special attention to cardinal numbers see also Freguglia (1982).

2 The explicit use of notions such as reflexivity, symmetry and transitivity in the Peano school seems to originate with Vailati (1892) and De Amicis (1892). Vailati in 1892 claims originality for introducing the word 'reflexivity'. De Amicis also credits Vailati with the introduction of 'reflexivity' and both credit de Morgan with the introduction of 'transitivity'. De Amicis coined 'convertible' [conversivo] for what we call 'symmetric' but his terminology did not catch on. Symmetric, in this sense, was introduced by Schröder in 1890.

3 This is the only time that a number-theoretic example is given in this context by the Peano school. This occurrence does not contradict my claim in Mancosu (2015a) that number-theoretic examples were not considered candidates for abstraction by Frege and the Peano school. In a

3. the straight line a is parallel to the straight line b;
4. the straight line a coincides with the straight line b;
5. figure a can be superposed over figure b.

These are called equalities according to the definition given in section 1. The relation of identity is a case of equality but not every equality coincides with identity. Peano explains that one can define several equalities over a system of entities depending on the specific properties of the entities one decides to take into consideration.

> Every equality between the entities of a system that is different from identity is equivalent to the identity between the entities that are obtained from those of the given system abstracting from all and only those properties that distinguish an entity from those equal to it. Thus, the equality "the segment AB can be superposed over the segment $A'B'$" is equivalent to the identity between entities that can be obtained from every segment by abstracting from all those properties that distinguish it from all those to which it can be superposed. The entity that results from this abstraction is called the *magnitude* [grandezza] of the segment; the former equality is thus equivalent to the identity of the magnitudes of the two segments. If we agree to indicate identity with the sign =, the equality just considered can be written as
>
> $$grAB = grA'B'$$
>
> Analogously, the equality "the line AB is parallel to the line $A'B'$" can be written
>
> $$\text{direction } AB = \text{direction } A'B'$$
>
> and so on.[4] (Peano (1888), pp. 152–154)

nutshell, the reason is that number theory in the nineteenth century worked with representatives of the equivalence classes and thus the function associated to the equivalence relation yields not new abstract elements but rather elements of the domain over which the equivalence relation is defined (so, numbers (minimal residues) in the case of congruence and minimal quadratic forms in the case of quadratic forms).

4 Ogni uguaglianza tra gli enti di un sistema, diversa dall'identità, equivale all'identità tra gli enti che si ottengono da quelli del sistema dato astraendo da tutte e sole quelle proprietà che distinguono un ente dai suoi eguali. Così l'eguaglianza "il segmento AB è sovrapponibile al segmento $A'B'$" equivale all'identità tra gli enti che si ottengono da ogni segmento astraendo da tutte le proprietà che lo distinguono da quelli con cui è sovrapponibile. L'ente che risulta da questa astrazione viene chiamato grandezza del segmento; l'eguaglianza precedente equivale quindi all'identità delle grandezze dei due segmenti. Se conveniamo di indicare col segno = l'identità, l'eguaglianza ora considerata si potrà scrivere:

$$grAB = grA'B'$$

Analogamente, l'eguaglianza "la retta AB è parallela ad $A'B'$" si può scrivere:

$$\text{direzione } AB = \text{direzione } A'B'$$

e così via.

Peano comes back to the topic of definitions by abstraction in 1894 in his "Notations de Logique Mathématique". After having discussed explicit definitions, in §38 he turns to a new sort of definition that is important in mathematics (there is no reference to Frege in Peano's discussion). He actually does not quite talk of 'definition by abstraction' but he claims that one introduces concepts by abstraction and by doing so one defines an equality.

> There are concepts that are obtained by abstraction which constantly enrich the mathematical sciences but that cannot be defined in the stated form [namely, with an explicit definition, PM]. Let u be an object; by abstraction one obtains a new object φu; one cannot form an equality
>
> $$\varphi u = \text{known expression}$$
>
> since φu is an object whose nature is completely different from all those that have hitherto been considered. Hence one defines the equality by stating
>
> $$h_{u,v} . \rightarrow : \varphi u = \varphi v . = . p_{u,v} \qquad \text{Def.}$$
>
> where $h_{u,v}$ is the assumption on the objects u and v; $\varphi u = \varphi v$ is the equality that is being defined; it has the same meaning as $p_{u,v}$, which is a condition, or relation, between u and v,[5] with a well known meaning. (Peano (1894), p. 95)[6]

The claim that the left-hand side and the right-hand side of the equivalence have the same meaning will be repeated by many in the Peano school but it is of course something many people will dispute.[7]

Then Peano states that the equality among the newly introduced objects, and a fortiori the equivalence relation, must satisfy the properties of reflexivity,

5 For typographical reasons, I have replaced throughout the Peano symbol for material conditional with $\rightarrow$.

6 "Il y a des idées qu'on obtient par abstraction, et dont s'enrichissent incessamment les sciences mathématiques, qu'on ne peut pas définir sous la forme énoncée. Soit u un objet; par abstraction on déduit un nouveau objet φu; on ne peut pas former une égalité

$$\varphi u = \text{expression connue},$$

car φu est un objet de nature différente de tous ceux qu'on a jusqu'à présent considérés. Alors on définit l'égalité, et l'on pose

$$h_{u,v} . \rightarrow : \varphi u = \varphi v . = . p_{u,v} \qquad \text{Def.}$$

où $h_{u,v}$ est l'hypothése sur les objets u et v; $\varphi u = \varphi v$ est l'égalité qu'on définit; elle signifie la même chose que $p_{u,v}$ qui est une condition, ou relation, entre u et v, ayant une signification bien connue." Frege makes parallel comments in *Grundlagen*, § 69; see Brandom (1986).

7 It is, for instance, denied by Crispin Wright and Bob Hale who however try to articulate in which sense the left-hand side and the right-hand side of the equivalence have the same 'content'.

symmetry, and transitivity (this time using this terminology; see also Burali-Forti (1894b)). I would like to draw attention to an important shift between this presentation and the one in 1888. One of the examples given by Peano in 1888 was that of congruence in number theory. In 1894 congruence is not mentioned among the examples. And this is certainly not on account of the fact that congruence in 1894 had ceased to satisfy the relevant properties for being an equivalence relation. Rather, the ontological spin given by Peano to definitions by abstraction, namely what is introduced is a new object φu, does not capture what happens in number theory in the nineteenth century where the object introduced is usually a representative of the equivalence class and thus not a new object (see footnote 4).

Peano says that a relation that satisfies the three properties in question has 'the properties of equality'. The object denoted by φu is what one obtains considering all and only the properties that it has in common with all other objects v that are equivalent to u, so that one also has φv.

Peano then provides a list of examples, beginning with the theory of ratios in Euclid. He also presents examples taken from arithmetic (broadly construed): the introduction of rationals from the integers by means of pairs using Stolz (1885), the introduction of irrationals as *lim sups* of sets of rationals. Moving to geometry, Peano mentions the introduction of length and direction. About the latter he writes:

> The relation between two unbounded straight lines "a is parallel to b" has the properties of equality. It has been transformed into "direction of a = direction of b", or "point at infinity of a = point at infinity of b". One cannot define an equality of the form: "point at infinity of a" = "expression formed with the words of Euclid's Elements" (Peano, 1894, p. 47)[8]

The latter remark is doubly connected to the ontological role that Peano ascribes to definitions by abstraction. It is exactly because the entity is undefinable using the previous vocabulary that a definition by abstraction results in something ontologically fruitful. But once again, I repeat, this also shows why the normal use of abstraction in number theory is, from this point of view, spurious in that the entity introduced (the representative) is simply one of the old entities and thus definable if the old entity was definable or available as a primitive entity. By way of further geometrical examples, Peano mentions

8 La relation entre deux droites illimitées "la a est parallèle à la b" a les propriétés de l'égalité. Elle a été transformée en "direction de a = direction de b", ou "point à l'infini de a = point à l'infini de b". On ne peut pas former une égalité de la forme: "point à l'infini de a" = "expression composée avec les mots des Éléments d'Euclide"

the introduction of vectors, quaternions as pairs of vectors, and concludes with Cantorian cardinalities, points beyond infinity in hyperbolic geometry, and Grassmann's geometrical forms. He also warns that the introduction of definition by abstraction is not always fruitful or desirable. He mentions as an unfruitful abstraction the case of 'shape' (arising from similarity of geometric figures) and then states that projectivity also satisfies the conditions for introducing an equality.

In section 40, Peano comments on the fact that the equalities introduced by a definition by abstraction are true identities and explains why there are no failures of substitutivity (i.e. why one cannot argue from $2/3 = 4/6$ and $2/3$ is an irreducible fraction to $4/6$ is an irreducible fraction).

At this point, many people in the Peano school began writing about definitions by abstraction. I will also recall that Russell in 1900 discovered Peano's contributions and this led to his (Fregean) techniques of eliminating definitions by abstraction in terms of explicit definitions. But such techniques had also already been anticipated by Burali-Forti, who will later repudiate them.

Let us consider Burali-Forti's *Logica Matematica* of 1894b.[9] Burali-Forti offers a taxonomy of definitions into four types and considers definitions by abstraction as the fourth type in his classification (pp. 140–145). The exposition is very similar to that given by Peano in 1894 although there are small variations in terminology. Burali-Forti explains that such definitions are used "when the entity x that one wants to define is obtained as an abstraction of a determined complex of known entities". The primary example discussed by Burali-Forti concerns the definitions of the abstract entities 'rational numbers'. Burali-Forti comments:

> In some cases not even the previous type of definition [by postulates, PM] can be adopted. This happens when the entity x that one wants to define is obtained as an abstraction of a determined complex of known entities.
>
> A rational number, for instance, can be defined as an abstract entity obtained from a pair of integers. If m, n are integers, with m/n we indicate an entity which depends on m and n and the mode of dependency, which can be stipulated at will (except for keeping in mind what result one wants to reach), defines, by abstraction, the entity m/n.
>
> In general: if u is the known entity (for instance, the pair m, n of integers), then the thing that one wants to define (for instance, the rational number m/n), is a function φ of u. For φu one needs to define the relation indicated by the sign $=$, saying what is the meaning, for the things φu, φv, of the relation $\varphi u = \varphi v$. Such Def has the form
>
> $$h_{u,v}. \rightarrow : \varphi u = \varphi v . = . p_{u,v}$$

9 Burali-Forti's *Logica Matematica*, in both the 1894 and 1919 editions, has been recently reprinted with an insightful introduction by Gabriele Lolli (see Burali-Forti (2013) and Lolli (2013)).

> where $h_{u,v}$ is the assumption relative to the things u, v; $p_{u,v}$ is the proposition, whose meaning is already known, containing u, v, and that we set equivalent to the relation $\varphi u = \varphi v$ that has to be defined.[10] (Burali-Forti (1894b), p. 140)[11]

However, Burali-Forti does not seem to realize that this only defines an identity between terms of the form $\varphi u = \varphi v$ and not an arbitrary identity $\varphi u = x$ (for an x which is not given in the form φv, for some v). Indeed, he goes on to assert that one can also provide an explicit definition of the class of entities:

> The class H of entities defined as abstract functions φ of known entities $u, v, \ldots$ is defined, by means of a definition of *first species* [i.e. an explicit definition, PM], by setting
>
> $$H = x\varepsilon(u\varepsilon M.x = \varphi u.-=_u \Lambda)$$
>
> where M is a known class, and this definition is read "H is the complex of entities x such that there exists at least a u in the class M for which it holds that x is *identical* to φu"[12] (Burali-Forti (1894b), p. 141)

Frege had given up definitions by abstraction because the equality $x = \varphi y$, where x is not given in the form φz, for some z, is not specified by the definition by abstraction.[13] One should add that in the Fregean case the situation was more

---

10 In certi casi neanche la forma precedente di definizione può essere adottata. Ciò avviene quando l'ente x che si vuole definire, si ottiene come astrazione di un complesso determinato di enti noti. Un razionale, p. es., può esser definito come ente astratto ottenuto da una coppia di numeri interi. Essendo m, n due numeri interi, con m/n indichiamo un ente che dipende da m e da n, e il modo di dipendenza, che noi possiamo stabilire ad arbitrio, (salvo il risultato al quale si vuol giungere), definisce, per astrazione, l'ente m/n. In generale: se u è la cosa nota (p. es., la coppia m, n di numeri interi), allora la cosa che si vuol definire (p. es. il razionale m/n), è una funzione φ di u. Per φu bisogna definire la relazione indicata dal segno $=$, dicendo quale è per le cose φu, φv, il significato della relazione $\varphi u = \varphi v$. Tale Def ha la forma

$$h_{u,v}. \rightarrow : \varphi u = \varphi v . = . p_{u,v}$$

ove $h_{u,v}$ è l'ipotesi relativa alle cose u, v; $p_{u,v}$ è la prop. contenente u, v avente già significato noto, e che poniamo equivalente alla relazione $\varphi u = \varphi v$ da definire.

11 Burali-Forti immediately goes on to mention that, starting from an appropriate relation, one can also define new abstractions on abstracts entities.

12 La classe H di enti, definiti come funzioni astratte φ degli enti noti $u, v, \ldots$ risulta, con una definizione di *prima specie*, definita, ponendo

$$H = x\varepsilon(u\varepsilon M.x = \varphi u.-=_u \Lambda)$$

ove M è una classe nota, e tale definizione si legge "H è il complesso degli enti x tali, che esiste almeno un u della classe M, per il quale x è *identico* a φu".

13 On account of one of the examples used by Frege in his discussion this has become known as the Caesar problem.

dire on account of the need to use such equalities in contexts in which they simply could not be eliminated.[14]

Burali-Forti discusses in detail the introduction of rational and irrational numbers. He points out that the right-hand side in such definitions must be given by a relation which satisfies the properties of reflexivity, symmetry, and transitivity, – properties that are in turn inherited by the equality so that $\varphi u = \varphi u$, if $\varphi u = \varphi v$ then $\varphi v = \varphi u$, and if $\varphi u = \varphi v$ and $\varphi v = \varphi w$ then $\varphi u = \varphi w$. This is followed by some geometrical examples with equivalence of plane figures, parallelism of lines and planes (thereby introducing areas, directions or points at infinity, and orientation [*giacitura*]). I remark that the relation of congruence modulo a certain natural number is not given as an example (nor is any other example from number theory). Finally, an interesting example comes with the notion of 'meaning' [*significato*]. Burali Forti abstracts from logically equivalent propositions to obtain the 'meaning' or the 'value':[15]

> If A and B are propositions, the relation $A \rightarrow B.B \rightarrow A$ or, A is *equivalent* to B, is reflexive, symmetric, and transitive (p. 27). We obtain then from each proposition A the abstract entity *value of* A or meaning of A: and we say that "The meaning of A is *equal* to the meaning of B, just in case A is equivalent to B". An analogous observation holds when A and B are classes.[16](Burali-Forti (1894b), p. 147)

When applied to classes, this principle is nothing else than Frege's Basic Law V (whether Burali-Forti had seen Frege's 1893 work at this stage, I do not know; Peano published a review of it in 1895). Finally, Burali-Forti explains that there is no failure of substitutivity originating from the definitions by abstraction under consideration. The examples are the same as those Peano used in 1894 concerning irreducible fractions.[17]

---

14 The issue, which concerns Frege's definition of the successor through the ancestral, has been discussed at length in Dummett (1991), [Hale and Wright 2001], and [Heck 2011], just to name three prominent examples from the extensive literature on this matter.

15 (Brandom (1986), p. 281 and p. 292) persuasively points out that all semantical notions in Frege (sense, reference, thought, truth value etc.) are introduced by means of definitions by abstraction.

16 Se A, B sono proposizioni, la relazione $A \rightarrow B.B \rightarrow A$ o, A è *equivalente* a B, è riflessiva, simmetrica e transitiva (p. 27). Otteniamo allora da ogni prop. A l'ente astratto *valore* di A o significato di A: e diciamo che "Il significato di A è *eguale* al significato di B, quando A è equivalente a B". Analoga osservazione vale quando A, B sono classi.

17 Così, p. es., 4/5 è *frazione irreduttibile*; ma 4/5 = 8/10; dunque 8/10 è *frazione irreduttibile*. Il che è falso. Ed è falso perchè *frazione irreduttibile*, non è una funzione del razionale 4/5, ma invece una funzione della coppia (4, 5), diversa dalla funzione $R^{\cdot}(4, 5)$ o 4/5. Se dunque definiamo il razionale come una funzione di una coppia di numeri, con i termini *frazione*

A perusal of Burali-Forti's work in set theory during this period shows that the terminology preferred by him is that of 'introduction of abstract entities'. Thus in 1896a in his work on finite classes, he presents the Cantorian introduction of cardinal numbers by means of a definition by abstraction and claims that the left-hand side and the right-hand side of the equivalence have the same meaning. The finite cardinalities are seen as 'abstract entities' which are the values of functions with domain the finite classes (see note 1, p. 51).[18] In his 1896b, Cantorian cardinalities and ordinal numbers are also introduced by means of definitions by abstraction but without using this terminology and talking about the introduction of 'abstract entities' which are functions of given classes or, as Burali-Forti says in 1894a, p. 177, one can 'define a class of abstract entities' (ordinal numbers) for which 'the identity is defined' by the appropriate equivalence.

In this 1894a paper, Burali-Forti refers to Peano (1894) which shows Peano's leading role in these reflections (we have already seen that Peano had written about such matters already in 1888). As far as I have been able to establish, the first complete and explicit use of the expression 'definition by abstraction' appears in a review of Burali-Forti (1894b) written by Vailati in 1894. In the review, Vailati says that "the Author [Burali-Forti] gives special attention to the particular case of the so-called definition by abstraction [*definizione per astrazione*]". From now on the terminology becomes standard. Peano will use it in several articles and contributions starting in 1899 (see Peano (1899a), p. 12; Peano (1899b), p. 135; Peano (1900), p. 13, and then in many publications and reviews in 1901). Especially interesting is the definition given in the dictionary of mathematics where he shows awareness that a definition by abstraction does not define in isolation the terms flanking the equality on the left-hand side of the biconditional. He says:

> Abstraction. In mathematical logic one calls "definition by abstraction" the definition of a function φx with the form: $\varphi x = \varphi y. = .$ (expression formed with previously given signs), that is one does not define the sign φx in isolation, but only the equality $\varphi x = \varphi y$.[19] (Peano (1901c), p. 7)

irreduttibile non indichiamo più un razionale nel senso inteso prima." (Burali-Forti (1894b), p. 148)

18 Incidentally, a detailed study of this article by Burali-Forti would repay detailed attention. He aims at proving the Peano axioms for arithmetic from the abstraction principle for finite Cantorian cardinalities using only the logical notions of class and correspondence.

19 Astrazione. Dicesi in logica matematica "definizione per astrazione" la definizione di una funzione φx avente la forma: $\varphi x = \varphi y. = .$ (espressione composta coi segni precedenti), cioè non si definisce il segno isolato φx, ma solo l'uguaglianza $\varphi x = \varphi y$.

In order to complete the treatment of the pre-Russell phase, let us consider briefly Burali-Forti (1899), Burali-Forti (1901) and Peano (1901a), Peano (1901b).

Burali-Forti (1899) shows already from its title the preoccupation with a characterization of equality, which is carried out in terms of the by now familiar technique of introduction of new abstract objects through a function operating on elements that are related by an equivalence relation. Using the standard stock of examples (length, area, volume, shape, direction, orientation etc.) from geometry (once again, congruency modulo n is not used), Burali-Forti recaps the elements of the theory and the distinction between equality and identity. The article is of interest only because it contains the roots of an 'operator' approach to the introduction of new entities, such as the rational numbers, that Burali-Forti will in the following years try to present as an alternative way to introduce by an explicit definition the entities normally obtained through a definition by abstraction. Indeed, he claimed that his definitions of rational and irrational numbers are to be considered as explicit definitions if the notions of magnitude and correspondence are granted.[20] Burali-Forti went on to claim that it is only when one cannot give such an explicit definition that another approach is necessary and lists the definitions of cardinal number, ordinal number, direction, orientation, length, areas, volumes, mass, and temperature as relevant cases. But rather than describing the usual definitions by abstraction, he proceeded to show how in effect one can obtain any definition by abstraction as a consequence of an explicit definition of the class of elements that constitutes the range of the abstracting function. The text is hard to parse because rather than defining the range through an equivalence relation on which the abstraction must be considered, he defines the range by means of the function itself.

Moving now to the essential cases of definition by abstraction, Burali-Forti explains that by saying that x is a length one claims that x is a simple element and spells out this condition by stating that for such an element the phrase 'y is an x' does not have any meaning. He articulates a linguistic distinction between *length* and *length of*. Suppose length(x) ('length' is not italicized) is the function given by the definition by abstraction. For a specific segment a, length(a) is a simple element called the length of a. Then Burali-Forti explains:

> When we say, for instance, that x is a length, we mean to say that x is a *simple element* that is a function (for example) of a segment [this simply means that x = length(a) for a segment a, PM], and this function [that is length(a), which is x] is common to all the segments that can

20 The same type of 'operator' approach is defended, in contraposition to the introduction of rationals by abstraction, in Peano's 1901a. For some considerations on Burali-Forti's operator theory, which I cannot further discuss here, see Lolli (2013).

> be superimposed to each other. If a is a segment, the function of a being considered [namely length(a), that is x] is indicated by the phrase *length of a*. Thus, while the word length is a **common noun**, a **class**, the word *length of* is a **correspondence** between the *segments* and the class *length*. In the same way, *cardinal number* indicates a class and *cardinal number of* indicates a correspondence between the classes and the simple elements that are the elements of the class *cardinal number*.[21] (Burali-Forti (1899), pp. 257–258)

Thus, *length* stands for a class containing the different lengths of segments, and thus it is not a simple entity, whereas each length(a), for a segment a, is a simple element which is a member of *length*. In the case of cardinal numbers *Cardinal* is the class, thus a complex entity, of all simple elements that are values of functions preserving equinumerosity. I will not get into the details of the kind of advantage Burali-Forti thought he had accomplished by making this move. He defined the notion of cardinal number as 'one of the correspondences f between classes and simple elements such that for any class u, the classes v for which $fv = fu$ are all, and the only, classes similar to u'. In his discussion he also realized that this definition was not unique on account of the fact that several functions could satisfy the relevant condition. The most important thing concerning this contribution is that Burali-Forti thinks of the individual cardinals (as opposed to the class *Cardinal*) as simple elements that cannot be further analyzed. This is in contrast to Russell's definition of each cardinal as a class of classes.

In 1900, Russell met Peano and some members of his school in Paris and, as is well known, this was a turning point in his intellectual career. At the meeting in Paris, Peano had talked about definitions and Burali-Forti, who could not be present, sent a contribution on the definitions of irrational numbers that was read by Couturat. Peano (1901a) has nothing to offer[22] on the matter of definition by abstraction but Burali-Forti's article (1901) goes back to the issue

---

21 Lorsque nous disons, par example, que x est une longueur, nous entendons dire que x est un *element simple* qui est fonction (par exemple) d'un segment [this simply means that x = length(a) for a segment a], et cette fonction [that is length(a), which is x] est commune à tous les segments superposables entre eux. Si a est un segment, la fonction considérée de a [namely length(a), that is x] est indiquée par la phrase *longueur de a*. Donc, tandis que le mot *longeur* est un **nom commun**, une **classe**, le mot *longueur de* est une **correspondence** entre les *segments* et la classe *longueur*. De même, *nombre cardinal* indique une classe, *nombre cardinal des* indique une correspondence entre les *classes* et des éléments simples qui sont les elements de la classe *nombre cardinal*.

22 Actually, in this paper (but this was added only in 1901 after the lecture was delivered), Peano contrasts in a footnote (p. 286) the introduction of fractions by abstraction and the 'operator' approach he is championing.

and is relevant for us. Burali-Forti claims to be able to classify all definitions into three sorts: nominal, by postulates, and by abstraction. From the classification he attempted to draw an important philosophical distinction between concepts and intuitions. His idea was that any x that is defined by a nominal (explicit) definition is a concept. Intuitions are those x's that can be given only through a definition by postulation or a definition by abstraction. It follows, according to Burali-Forti, that whether something is a concept is an absolute notion, whereas whether something is an intuition depends on the state of science. Indeed, notions that were introduced by postulation (as in axiomatic systems) or by abstraction might then be able to be explicitly defined at a later stage of research. The formal characterization of definition by abstraction given by Burali-Forti in this article is standard but he repeats that one can define nominally the range of an abstraction function. For instance *direction* is $\{x$: there is a straight line a such that $x = direction\ of\ a\}$. Notice that, as previously explained, there is a difference between *direction* and *direction of*. The former is a class made up of simple entities (the directions of a, for arbitrary segments a). On the basis of this opposition between concepts and intuitions, Burali-Forti classified the definitions of number given by Dedekind and Peano (by postulates) and by Cantor (by abstraction) all as intuitions. He then contrasted those definitions with his definition of natural number, which he claimed to be an explicit definition. He made the same claim for his theory of rational numbers and the integers. I will not enter the details of the alleged explicit definition given by Burali-Forti for it would force me to present his theory of magnitude. I will only mention that it is clear that we have here some sort of foundational program that aims at eliminating definitions by postulation and definitions by abstraction in favor of nominal (explicit) definitions. I think this goal was also shared by Peano and affects Russell's approach to this issue. Regardless whether one can grant Burali-Forti the successes he claimed, both Peano and Burali-Forti agreed that Cantor's theory of cardinal and ordinal numbers could not be reduced to explicit definitions at the then current state of science.

2 Russell and Couturat

The above was the state of the discussion on definition by abstraction when Russell discovered the Peano school and entered the fray with his article on relations, Russell (1901). Not surprisingly, we find Russell emphasizing the issue of when a definition by abstraction can be turned into a nominal definition. There is a draft of Russell (1901) published in vol. 3 of the *Collected*

Papers[23] where Russell's first reference is to Burali-Forti (1901) (Russell (1993), p. 590). Even the first claim, in the original English version, about the advantage of introducing functions defining them through the notion of relation is strictly connected to the previous debate:

> The following notation appears to introduce at once a simplification and a generalization of many mathematical theories; and it enables us to render all definitions nominal. (Russell (1993), p. 590)[24]

Russell's contribution to abstraction have already been the subject of scholarly scrutiny and thus I will be brief.[25] Gregory Moore summarizes quite clearly the major results of the paper on relation vis-à-vis definition by abstraction:

> To obtain a definition of cardinal numbers, he [Russell] used what Peano called "definition by abstraction". That is, given an equivalence relation R, there is a function φx such that xRy if and only if $\varphi x = \varphi y$; thus, for example, the relation of one-one correspondence between two classes x and y gives rise to the function "cardinal number of x" (Peano (1894), p. 45). But Russell regarded it as necessary, in order to obtain such a function φ (or, as he preferred to put it, a many-one relation S), to introduce a primitive proposition stating that any equivalence relation R can be written as the relative product of a many-one relation S and its converse (V.I, §I, *6.2). He applied this primitive proposition, which in the *Principles* he called the Principle of Abstraction (1903, p. 166), to the relation of similarity between classes, thereby obtaining such a relation S; then he defined the class Nc of all cardinal numbers as the codomain of S. But in a marginal comment by this definition, he recognized the problem that S is not uniquely determined: "This won't do: there may be many such relations as S. Nc must be indefinable" (V.1, §3, *1.4). While in his first draft (Appendix V.I) he freely referred to the cardinal number of a class without any mention of S, in the published version (Paper 8) he took a particular S as given and only defined individual cardinal numbers in terms of S. Nevertheless, sometime between February and July 1901, he added a sentence to the effect that, for any equivalence relation R, we can always take the equivalence class of a term u as "the individual indicated by the definition by abstraction; thus for example the cardinal number of a class u would be the class of classes similar to u" (8, 320). This was the famous Frege-Russell definition of cardinal number. Russell applied it not only to the relation of one-one correspondence, in order to obtain cardinal numbers, but to any equivalence relation whatever. (Russell (1993), p. xxvii)

---

23 *Caveat lector*: the English translation published by Marsh and corrected by Russell in 1956 has important changes with respect to the original, in particular on the issues I am discussing.

24 The 1956 translation reads like the French published text: "and it permits us to give nominal definitions whenever definitions are possible". (Russell (1993), p. 315) The reference to Burali-Forti, which was in the English draft (Russell (1993), p. 590), was however removed from the published version.

25 See, for instance, the analysis of Russell on definitions by abstraction given in Vuillemin (1971), Rodriguez-Consuegra (1991), and Grattan-Guinness (2000).

In Russell's own words:

> P6.2 is the converse of P6.1. It affirms that all relations which are transitive, symmetrical, and non-null can be analyzed as products of a many one relation and its converse, and the demonstration gives a way in which we are able to do this, without proving that there are not other ways of doing it. P6.2 is presupposed in the definitions by abstraction, and it shows that in general these definitions do not give a single individual but a class, since the class of relations S is not in general an element. For each relation S of this class, and for all terms x of R, there is an individual that the definition by abstraction indicates; but the other relations S of that class do not in general give the same individual. [...] Meanwhile, we can always take the class ρx, which appears in the definition of Prop 6.2, as the individual indicated by the definitions by abstraction; thus for example the cardinal number of a class u will be the class of classes similar to u. (Russell (1993), p. 320; [preliminary English draft of Russell (1901)])

According to Russell then, we obtain in this way a nominal definition of the cardinality of a class (in addition to the concept of cardinality itself). Indeed, in *Principles of Mathematics* (1903, p. 112) he will argue that the new theory of relations is the only one that allows to give up both definitions by postulation and by abstraction. Once again, he refers in a note to Burali-Forti (1901):

> Moreover, of the three kinds of definitions admitted by Peano — the nominal definition, the definition by postulates, and the definition by abstraction [a note here refers to Burali-Forti (1901)] — I recognize only the nominal: the others, it would seem, are only necessitated by Peano's refusal to regard relations as part of the fundamental apparatus of logic, and by his somewhat undue haste in regarding as an individual what is really a class. (Russell (1903), p. 112)

Like Burali-Forti, Russell worries about the non-uniqueness of the relation S (Burali-Forti's worry was the analogous one about the function corresponding to S). These latter issues will be central in the later discussion on Russell's transformation of definitions by abstraction into nominal definitions.

The treatment of abstraction in *Principles* is somewhat confused due to the stratified nature of the composition of the text. In the middle of writing the book, Russell discovered his technique for eliminating abstraction but some passages in the text reflect an older state of things. This much he admitted in a letter, dated December 10, 1903, addressed to Couturat who had asked for clarifications on this matter (Russell (1993), pp. 726–727; Couturat's letter was dated December 7, 1903, see Schmid (2001), pp. 343–345).

In *Principles*, Russell proceeded, almost in Fregean manner, by first offering a definition of number by abstraction only to criticize it and replace it with a

nominal definition.[26] Unlike Burali-Forti, Russell takes the values of the abstraction function to be classes and thus not simple entities. Peano in 1901b, p. 70, explicitly provides a definition by abstraction of Cantorian cardinalities, using the symbolism $Num(a)$ for a class a, but explicitly rejects the idea of identifying $Num(a)$ with the class of all classes that are in one-one correspondence with a, for he explains that $Num(a)$ and the classes of classes that are in one-one correspondence with it are "objects which have different properties." It is unclear whether Peano's refusal to identify $Num(a)$ with a class originates from a belief, shared with Burali-Forti, that $Num(a)$ must be simple or whether purity of methods issues might be at stake.[27] I will recall here that when Dedekind discussed a proposal by Weber concerning a theory of number in which the numbers are defined as classes of classes, he rejected the proposal exactly with motivations very similar to those of Peano's (see Reck (2003), p. 385). Dedekind wrote:

> If one wishes to pursue your approach I should advise not to take the class itself (the system of mutually similar systems) as the number (Anzahl, cardinal number), but rather something new (corresponding to this class), something the mind creates. (cited in Reck (2003), p. 385)

Russell rejected Peano's position and claimed not to be able to see what these different properties between the two entities might be. He added:

> Probably it appeared to him immediately evident that a number is not a class of classes. But something may be said to mitigate the appearance of paradox in this view. (Russell (1903), p. 115)

Russell adduced some considerations for making it less objectionable to identify numbers with classes and concluded that the strategy outlined for cardinal numbers could be used in all definitions by abstraction:

> Wherever Mathematics derives a common property from a reflexive, symmetrical and transitive relation, all mathematical purposes of the supposed common property are completely served when it is replaced by the class of terms having the given relation to the given term; and this is precisely the case presented by cardinal numbers. For the future, therefore, I shall adhere to the above definition, since it is at once precise and adequate to all mathematical uses. (Russell (1903), p. 116)

---

26 I shall first set forth the definition of numbers by abstraction; I shall then point out formal defects in its definition, and replace it by a nominal definition. (Russell (1903), p. 112)

27 Such objections are repeated in Burali-Forti (1909) and Catania (1911). According to Burali-Forti a "simple entity is one that is not a class." For Peano's alternative conceptions of the numerosity function see Mancosu (2015b).

Russell's main objection to definitions by abstraction consists in the fact that the function postulated in the abstraction is not unique:

> Now this definition by abstraction, and generally the process employed in such definitions, suffers from an absolutely fatal formal defect: it does not show that only one object satisfies the definition. Thus instead of obtaining *one* common property of similar classes, which is *the* number of the classes in question, we obtain a class of such properties, with no means of deciding how many terms this class contains. (Russell (1903), p. 114)

In section 111 of *Principles* two possible solutions to the lack of uniqueness were discussed and eventually Russell settled for an identification of the number of a class a with the class of all classes that are in one-one correspondence [similar] to a. As we have seen, Russell proposed to apply this strategy to all definitions by abstraction.

Couturat (1905) also rejects definitions by abstraction in favor of explicit definitions obtained through Russell's principle of abstraction. After having explained what this principle amounts to, Couturat (1905, p. 50, note 1) uses a turn of phrase that stems from Russell's letter to Couturat from December 10, 1903: "Thus, the principle of abstraction does not lead to an abstraction but on the contrary it allows one to dispense with abstraction and to replace it"[28]

---

28 "Le principe d'abstraction n'a donc pas pour resultat d'effectuer l'abstraction, mais au contraire d'en dispenser et de la remplacer." (Couturat (1905), p. 50, note 1) Since the exchange between Russell and Couturat is not easily available I add the part of the correspondence that is relevant here. Couturat to Russell, December 7, 1903: "Seulement, je voudrais avoir un eclaircissement sur le *principe d'abstraction*. Vous dites (p. 166) que vous avez appliqué ce principe aux nombres cardinaux. Or dans la 2e partie je ne vois pas que vous ayez fait usage de ce principe, puisque vous y définissez le nombre cardinal comme une classe de classes (p. 115, 136). Et pourtant vous dites, dans votre *Préface* (p. IX), que le principe d'abstraction vous permet de définir les nombres comme classes. C'est ce que je ne comprends pas. Vous n'avez pas besoin de ce principe pour définir par ex. les classes *équivalentes* (*similar*); et ce principe peut vous server à déduire d'une classe de classes équivalentes l'idée du nombre cardinal qui est leur propriété commune. Il vous fournit donc les nombre cardinaux comme des entités singulières, et non comme des classes de classes. De même, quand vous l'appliquez à des *quantités égales*, il vous fournit la grandeur commune à toutes ces quantités, c. à d. *une* et *identique* en toutes. Il semble que vous ayez oublié d'appliquer votre principe au nombre, car vous ne le formulez explicitement que dans la 3e Partie (p. 166, 220) ce qui est un peu tard. Autre question: Quelle est la valeur et la nature de ce principe? Ce n'est pas, apparemment, un principe premier, un axiome indémontrable (Pp.). Mais alors, comment le démontrez-vous? Cela me paraît difficile, car il est éminemment *synthétique*; en effet, il fait surgir d'une simple relation entre 2 entités une 3e entité nouvelle. Ecrivons en symbols:
$aSb. \rightarrow .aRc.bRc$
(S sym. et transitive; R rel. uniforme [many one]).

(see also Russell (1914)). Needless to say, this method of elimination is also followed by Whitehead and Russell in *Principia Mathematica* even though the set up is much more complicated on account of the predicative theory of types motivated by the paradoxes.

Indeed, all would have been logically unobjectionable had it not been for the discovery of Russell's paradox. The discovery, which took place in summer 1901, i.e. in the middle of writing *Principles*, affected the reduction of definition by abstraction at least in those cases, such as cardinal (see section 111 of *Principles*) and ordinal (see section 231 of *Principles*) numbers, where the classes (of classes) turned out to be paradoxical. Russell already discussed this problem in *Principles*. On p. 305 he admitted that his method of turning definitions by abstraction into nominal definitions "is philosophically subject to the doubt resulting from the contradiction set forth in Part I, ch.x". In this connection, Russell referred to the appendix of *Principles* where he discussed at length Frege's system and the contradiction he had obtained from the axioms postulated by Frege.

In light of the paradoxes it was not at all clear whether definitions by abstraction could be dispensed with and, in the case of set theory, Weber [1906], Weyl [1910], and Hausdorff [1914] were resigned to treat ordinal and cardinal numbers as objects introduced by abstraction about whose nature nothing more precise could be said.

The story of the various proposals addressing this paradox and the consequences for abstraction principles would have to take on board most of the debate on the foundations of mathematics between *Principles* and the formalization of most systems of logic and set theory well into the thirties. This is obviously not feasible here and the last section will instead look at the reactions within the

---

On comprend la déduction inverse:
$aRc.bRc. \rightarrow .aSb$ (car $S = R\breve{R}$)
qui *élimine* c, mais la déduction directe, qui *introduit* c (et même le determine) paraît un peu forte, c. à d. paradoxale." (Schmid (2001), pp. 343–344)

Russell replied as follows on December 10, 1903: "Au sujet du principe d'abstraction il se trouve sans doute une obscurité dans mon livre, qui resulterait de ce que je l'ai accepté autrefois comme axiome, tandis que je l'ai pu démontrer plus tard. La demonstration se trouve dans la Revue de Mathématique Vol. VII: je crois que le numéro est 6.28, mais je n'ai pas le volume ici. L'essentiel du principe, tel qu'il se démontre, est de substituer la classe même des objets dont il est question à la qualitè hypothétique commune à tous ces objets. Au lieu de "principe d'abstraction" j'aurais mieux fait de l'appeler "principe remplaçant l'abstraction". Quand on a une relation S symmétrique et transitive, la classe des objets qui ont avec a la relation S remplace, dans le calcul, la propriété commune à tous ces objets, que suppose le sens commun. Je ne nie pas qu'il y ait souvent une telle propriété, mais il n'est pas nécessaire de l'introduire; elle serait en général indéfinissable, et la classe a toutes les qualités dont on a besoin." (Schmid (2001), p. 346)

Peano school to the Russellian solution and to the problem of accounting for definitions by abstraction.

3 Padoa on definitions by abstraction and further developments

One of the best papers written on definitions by abstraction is Padoa (1908). Padoa starts with explicit definitions (which is his preferred terminology for what Peano and Burali-Forti called nominal definitions). In an explicit definition one sets a (definitional) equality between expressions meant to indicate that the new expression is introduced as short hand for a longer known expression. In this way the definiens acquires a meaning and the equality can be interpreted as 'means'. He then considered definitions by abstraction given in the following form. For K a known class and R an equivalence relation, the form of any definition by abstraction is:

1. If "a and c are (individuals belonging to) K" then "$Fa = Fb$" means "aRb"

where F is the function one wants to define. For instance: if "a and b are polygons", then "area of a = area of b" means a is equivalent to b. By keeping K as the class of polygons and letting R denote similarity, respectively equivalence, one defines the functions 'form of' and 'area of'. He added:

> Since the *defined* notation is "$Fa = Fb$", equation 1 does not *explicitly* indicate either the meaning of F or that of "Fa". In other words, it only *authorizes* the use henceforth of the *sole* expression "$Fa = Fb$", because (using the reverse procedure) it teaches to substitute only to this expression an expression of known meaning ("aRb").[29] (Padoa (1908), p. 94)

To begin with, Padoa entertains the possibility of considering that the definition by abstraction results in an explicit definition of the complex expression "$Fa = Fb$". He then asks whether this definition is arbitrary and points out that the question might appear paradoxical since all nominal definitions are arbitrary. Padoa argues as follows. In an explicit definition, say α, we have as a part of α the definiendum that contains new vocabulary with respect to the language available

29 Poichè la notazione *definita* è "$Fa = Fb$" la 1 non indica *esplicitamente* il significato nè di F nè di "Fa". In altri termini, essa *autorizza* ad usar poi la funzione F nella *sola* scrittura "$Fa = Fb$", perchè (usando il procedimento inverso) solo a questa essa insegna a sostituire una scrittura di significato noto ("aRb").

previously to introducing α. It would seem reasonable therefore that, using the old language and theory available before the introduction of the explicit definition α, α cannot be deduced from the theory. Padoa then rhetorically asks: would it not be reasonable also to say that α cannot be contradicted in the theory available before its introduction? But, he claims, this is not so, when the defined notion expresses a condition whose principal symbols are already known.

In order to explain his point, Padoa focuses on a relation R defined on objects of a class K. Padoa calls a relation R 'egualiforme' (I will translate this as 'equiform') if it is reflexive, symmetric and transitive.[30] Let me remark that this is the first time that equivalence relations receive a special name to characterize them. Now Padoa proves Theorem I, namely from the definition by abstraction

(1) if a and b are in K, "$Fa = Fb$" means "aRb", one infers
(2) R is 'equiform' in K

The argument, which relies on the assumption that the meaning of '=' on the left hand side of (1) is known, shows that (1) cannot be assumed arbitrarily unless (2) has already been established. In other words, the 'equiformity' of R is a necessary condition for introducing a definition such as (1). Thus, if from the theory previous to the introduction of (1), the negation of (2) were to result, (1) would lead to contradiction; by contrast, if the previous theory does not allow either to prove or to refute (2), then (1) plays both the role of a definition and of a postulate.

Having thus established that definitions such as (1) are not arbitrary, taken as an explicit definition of the complex "$Fa = Fb$" when R is assumed "equiform", leads Padoa to ask whether (1) might not be interpreted as an implicit definition of F. And here the main objection to using F only in contexts such as "$Fa = Fb$" does not originate, says Padoa, from purely formal considerations, but by the fact that (1) does not individuate the meaning [il significato] of F. Padoa gives many examples of this indeterminacy of meaning. Let's consider one of them. If a and b are polygons, aRb means "a is equivalent to b" (a and b can be decomposed in parts that are respectively superposable) and "area" is the intended F given in (1). He then remarks:

> We do not contest that (1) can be *legitimately* assumed as an *explicit Df* of "area of a = area of b"; but we deny that (1) *individuates* the *meaning* of "area of a".[31] (Padoa (1908), p. 97)

30 Even in the Italian contributions, this terminology will not assert itself. Burali-Forti [1912] speaks of a 'normal' relation and Cipolla [1914] of 'uniform' relation. Cipolla [1914] follows the account of definition by abstraction given by Russell and Padoa.

31 Non contestiamo che la (1) possa *legittimamente* assumersi quale *Df esplicita* di "area di a = area di b"; ma *neghiamo* che la (1) *individui* il *significato* di "area di a".

If the recipient had ignored the meaning of "*Fa*" as "area of", the definition by abstraction won't succeed in conveying it to him/her. In fact, one could attribute to *F* the meaning "twice the area of", "three times the area of", "half of the area of", etc. (where 'area' is still taken in the original informal meaning). In short "area of" as it appears in a definition of the form (1) can be interpreted with infinitely many different meanings. One interesting objection to this line of thought that Padoa considers consists in claiming that the meaning of "area of" given by (1) has a meaning that consists of "all the meaning compatible with (1)". And here Padoa's objection does not seem so strong, for he says:

> But then before (1) can teach us the meaning of "area of" one needs to have found *all* its meanings compatible with (1); and how can the reader verify to have considered *all* of them?[32] (Padoa (1908), p. 98)

How would, Padoa continues, the reader have included among the meanings of "area of *a*" that given by "volume of the prism having a base of area *a* and height a segment of constant length", which also satisfies (1)?

The point is repeated using examples with directions, orientations, etc. A useful footnote makes clear that in the case of directions one must use the relation "*a* is parallel to *b* or *a* coincides with *b*" which is an 'equiform relation' whereas "*a* is parallel to *b*" is not. This is connected to the fact that for much of the nineteenth century "*x* is parallel to *y*" was not taken to be reflexive.

So far the treatment has been critical. Now Padoa moves to a *pars costruens* and defines explicitly the "abstraction of *a* with respect to *R*" and proves a relevant theorem about it. Here is the passage:

> Let us stipulate first of all the following *explicit Df*: **Definition.** If *K* is a class and *R* is an *equiform relation* on *K* and if *a* is an arbitrary element of *K*, then "**abstraction** of *a* with respect to *R*" means "the set of all and only those *K* that bear the relation *R* to *a*." For instance: if *a* is a *polygon*, then "*abstraction* of *a* with respect to the relation *is equivalent to a*" is "the set of all and only those polygons *x* such that *x* is equivalent to *a*". Then we show the following **Theorem II**. If *K* is a *class* on which *R* defines an *equiform relation* and if *a* and *b* are in *K*, then (indicating, for the sake of brevity, with *Fa* and *Fb* the *abstractions* of *a* and *b* with respect to *R*): "*Fa* = *Fb*" if and only if "*aRb*".[33] (Padoa (1908), p. 100)

32 Ma allora, prima che la 1) ci apprenda il significato di "area di" bisogna aver trovato *tutti* i suoi significati compatibili con la 1); e come si accerta il lettore di averli considerati *tutti*?

33 Stabiliamo anzitutto la seguente *Df esplicita*: **Definizione.** Se *K* è una *classe* in cui *R* è una *relazione egualiforme* e se *a* è un individuo arbitrario di *K*, allora "**astrazione** di *a* rispetto ad *R*" significa "l'insieme di tutti e soli quei *K* che stanno nella relazione *R* con *a*." Ad es.: se *a* è un poligono, allora "astrazione di *a* rispetto alla relazione *è equivalente a*" è "l'insieme di tutti e soli i poligoni *x* tali che *x* è equivalente ad *a*". Poi dimostriamo il seguente

Of course, this is the Russellian solution and one wonders how Padoa, after the publication of Russell's *Principles* in 1903 and of Couturat (1905), could present this result without attributing it to Russell. Perhaps the answer is at the very beginning of the article where Padoa says that his reflections go back to 1901 and that what he had in mind to present at a conference in Livorno (at the Second Congress of the Italian Philosophical Society) at the time (1901) was later developed by Russell and Couturat. But since, he continued, their writings had not yet enjoyed a large readership and because they did not treat the matter with the simplicity and generality required, he had decided to write this article for the occasion of the meeting in 1906 from which the 1908 publication stems.

If theorem I showed that the relation R had to be 'equiform' as a necessary condition for the definition to be successful, theorem II provided a sufficient condition for it. Thus, the definition of "abstraction of" and theorem II constitute a "theory of abstraction". If one from the outset chooses, among the possible meanings for F, that entity given by "abstraction of", it turns out that a definition by abstraction of the form (1) is a theorem of the theory of abstraction. Coming back to the examples with areas of polygons we immediately reach an explicit definition of "area of a", for a polygon a, by setting "area of a" as "the abstraction of a with respect to the relation 'is equivalent to'". Then the original definition by abstraction can be regained as a theorem of the theory of abstraction. Incidentally, Padoa also makes no mention of Frege's work in this connection.

Padoa also remarked that one can simply get rid of the word "abstraction" and define "area of a" as "the set of polygons that are equivalent to a". We are now moving towards the standard set-theoretic territory. But Padoa does not seem to have yet digested the lesson of Russell's paradox, for he also applies the reasoning to defining the number of a finite class a as the class of all classes that are in one-one correspondence with a.

Padoa concluded with the following methodological lesson:

> One could object, I will not deny it, that common sense and experience will guide each time the writer to use definitions by abstraction exclusively when dealing with *equiform relations* on the considered class, thus avoiding any danger of contradiction.
>
> But it might have been useful to point out that if theorem I clarifies the *necessity* of such a *condition*, theorem II specifies its *sufficiency*.
>
> And although the aid, real or alleged, of intuition would lead each time the writer to consider the concepts *ambiguously* defined *by abstraction* (in the usual way) as perfectly

Teorema II. Se K è una *classe* in cui R è una *relazione egualiforme* e se a e b sono K, allora (per brevità, indicando ordinatamente con Fa ed Fb le astrazioni di a e di b rispetto ad R): "$Fa = Fb$" quando e sol quando "$a\,R\,b$".

> individuated, it will not have been useless to underline that the ambiguity was intrinsic to the *type* of *Df* [1] and to indicate a general procedure that — freeing us from the treacherous aid of *intuition* — will transfer the *explicit individuation* of *mathematical abstractions* from the psychological and epistemological domains to the *logical* realm.[34] (Padoa (1908), p. 103)

This last statement is a wonderful summary of the removal of the psychological roots of the abstraction process in favor of a merely logical description of it. It is exactly this move that bothered Angelelli (see 1984, 2004, 2013) but then again it seems to me that most of his worries reduce to an issue of semantics. Is it appropriate to refer to this definitional method with the word abstraction given that the tie with the psychological process has been severed? Perhaps not but what difference does it make? We could simply call the logico-mathematical process abstraction*.

On the contributions that follow this paper by Padoa, I will have to be very brief. An important text in this connection is *Elementi di Calcolo Vettoriale, con numerose applicazioni alla Geometria, alla Meccanica e alla Fisica-Matematica*, by Burali-Forti and Marcolongo (published in 1909). The text was translated into French in 1910. It is important for two reasons. The first is that Burali-Forti attempts to bypass the use of Russell's classes of classes (against which he argues in favor of considering the abstract entities as simple) and does so by stating a "logical postulate" to which the function F and the class that constitutes the range of this function are unique. This proposal was shown to be incoherent in Maccaferri [1913] and even though Burali [1912] already attempted to rectify his claim the only thing to point out about this latter article is Burali-Forti's rejection of definitions by abstraction in favor of a new operator theory that Burali-Forti will develop in the second edition of *Mathematical Logic* in 1919.

Incidentally, Maccaferri (1913) uses new examples to show the indeterminacy of meaning that is constitutive of definitions by abstraction but nothing he presents marks a decisive improvement on Padoa (1908), which Maccaferri dis-

34 Si potrà obbiettare, ed io non lo contesto, che il buon senso e l'esperienza guidano volta a volta il trattatista a giovarsi delle consuete *Df per astrazione* sol quando si tratti di *relazioni egualiformi* nella *classe* considerata, evitando così ogni pericolo di contraddizioni.
Ma può essere stato utile notare che, se il teorema I chiarisce la *necessità* di tale *condizione*, il teorema II ne precisa la *sufficienza*.
Ed ancorchè il soccorso, reale o presunto, dell'intuizione inducesse volta a volta il trattatista a ritener perfettamente *individuate* le idee *ambiguamente* definite *per astrazione* (nel modo consueto) — non sarà stato inutile rilevare che l'ambiguità — era insita a quel *tipo* di *Df* [1] e additare un procedimento generale che — affrancando dal malfido ausilio dell'*intuizione* — trasporti, dai campi psicologico e gnoseologico al campo *logico*, la *individuazione esplicita* delle *astrazioni matematiche*.

covered, as he declares in an afterword appended to the article, only after writing his article and with which he is pleased to agree. In his treatment Maccaferri ends up favoring the Russellian solution for turning definitions by abstraction into explicit definitions (or nominal definitions). He claims that nothing is simpler than choosing the Russell class as the value of the abstraction function and asks:

> Which function, of all the elements u related by the relation α, is *simplest* than the one that yields the very class of all those elements given that there is no criterion for choosing among them a unique element rather than any other one? And moreover every element b of the Russell class depends *only* on the elements that make it up, namely on the u's that are related among them by the relation α.[35]

The other reason why Burali-Forti and Marcolongo is of interest is for their lengthy appendix on definitions by abstraction that contains also the interesting definition by abstraction of Grassmann's forms. (See appendix to this article where the text is translated from the French edition).

Finally, I will mention Catania (1911) as an interesting discussion of the advantages and disadvantages of the Russellian principle of abstraction in comparison to that of Peano using only simple entities. Catania sides against Russell's approach and expresses optimism about the operator approach that Burali-Forti is developing (Catania himself will publish further work in this direction in the following years).

In Bindoni (1912) one finds an argument, resting on Burali-Forti's mistaken postulate, that the entities defined by abstraction and the class defined using nominal definitions (Russell-style) are identical. As for Peano (1911), it surprisingly contains no comments on definitions by abstraction. Peano (1915) however devotes a whole article to definitions by abstraction where he takes a rather pragmatic attitude as to which types of definitions are better. There are some interesting further examples of definition by abstraction relating to works by Fano on special relations between real numbers (such as 'belongs to the same algebraic field' or 'having a form $R \pm \sqrt{R}$') and different orders of infinities (Mago (1913)). Nothing Peano says in this article adds much to the previous discussion. However,

35 "Quale funzione, di tutti gli u legati dalla relazione α, *più semplice* che non sia quella che fa ottenere la classe stessa di quegli elementi, poichè non c'è un criterio per scegliere tra essi l'uno elemento piuttosto che l'altro? E d'altra parte ciascun elemento b della classe di Russell dipende *soltanto* dagli elementi che lo compongono, cioè dagli u legati fra loro dalla relazione α." (Maccaferri, 1913, p. 170)

he did claim credit for the expression 'definition par abstraction' by referring to his Peano (1894). But we have seen that this is only partially correct.

4 Conclusion

The foundational discussion on definition by abstraction goes on to include Burali-Forti (1919), Weyl (1910), Carnap (1929), Natucci (1923), Dubislav (1931) (3rd edition), and Scholz and Schweitzer (1935) (see also Cassina (1961)). The discussion in the 1920s and 1930s takes place against the background of the monumental *Principia Mathematica* and the use of definitions in context which is prominent in it. This naturally brings about a reconfiguration of the debate on definitions by abstraction but it is my sense that these works are mostly derivative from the previous discussion involving the Italian logicians centered around the Peano school, Russell, and Couturat.[36] In particular, the important ontological and semantical issues related to definitions by abstraction had already been characterized within the early discussion in the Italian school. Those discussions included:

(a) ontological issues (the nature of the entities obtained by abstraction, simple or complex?, and, if complex, can they be identified with classes?);
(b) semantical issues (indeterminacy of meaning and reference in the concept and terms, respectively, obtained by abstraction; sameness of meaning between the left hand side and the right hand side of the equivalence);
(c) logical issues (nature of the definition; necessary and sufficient properties required for its success; elimination of such definitions etc.)

Some of these points were of course foreshadowed by Frege and some were to reappear with a vengeance in the contemporary debate on neologicism.

5 Appendix

From Burali-Forti, C., and Marcolongo, R., *Eléments de calcul vectoriel avec de nombreuses applications à la géométrie, à la mécanique et à la physique mathématique*, Hermann, Paris, 1910, pp. 213–216.

36 Vuillemin (1971) contains many interesting developments including discussions of Whitehead, Carnap, Russell, etc.

Historical and Critical Notes

On definitions by abstraction

It is well known that, for instance, prime number and spherical surface can be given a definition (nominal or absolute) of the following form:

"prime number" = "integer with only two divisors"
"spherical surface" = "locus of points that are equidistant from a given point".

These are definitions in which the right hand side has a *well known* and *precise* meaning.

The same method of definition cannot be applied to *vector, geometrical forms, direction, length, weight, mass*, etc. For such entities it is better to employ a type of definition that can be called *definition by abstraction*. We deem it useful to state explicitly in what it consists and how one must apply it, for one often applies it in an unprecise fashion.

Equality

One can give, following Leibniz, an *absolute* and *universal* meaning to the sign =.

(1) "if x and y are arbitrary entities, one has $x = y$ if and only if *every property of x is also a property of y*"

And since "to be a property of x" can always be logically expressed by the statement "x is an element of a certain class u", the previous statement takes the form

(1′) "$x = y$ if and only if *every class u that contains x also contains y*."

Let us immediately make the following remark, which we will use later on: once we have defined in this fashion, following Leibniz, the relation designated by the symbol =, it is no longer allowed to define once more, for some newly introduced elements, the relation $x = y$. From (1) and (1') one infers:

(2) $\begin{cases} x = x \\ x = z \text{ and } y = z \text{ entails } x = y \text{ (Euclid)} \end{cases}$

and it is to these two properties that are reduced the *three* usual properties, reflexive ($x = x$), symmetric ($x = y$ entails $y = x$) and transitive ($x = y$ and $y = z$ entail $x = z$). It is customary to say that the properties given in (2) are the *characteristic*

properties of equality. This is not exact. In fact, the relations "it is equivalent to", "is similar to", for instance, satisfy the conditions in (2). However, they do not satisfy condition (1) because the relation "x is equivalent to y and y is a triangle" does not necessarily entail "x is a triangle". And similarly "x is similar to y and the volume of y is a cubic meter" does not necessarily entail "the volume of x is a cubic meter". In other words, condition (1) entails the conditions in (2) but the converse does not hold. While (1) defines a unique relation, *identity* or *absolute equality*, the properties in (2) define a class of *relations*, and the number of these relations is infinite, for we cannot put a bound to their construction. It follows that for the infinite relations defined by using (2), *it is not allowed to make use of the sign = if one wants to preserve this sign to designate, as it is customary, the precise Leibnizian sense of absolute equality or identity.*

Definition by Abstraction

This type of definition is based on the following principle of general logic.

Let us assume that for whichever elements x, y, z of a class u, the relation α between the [elements of] u, satisfies the following properties:

(3) $\begin{cases} x\alpha x \\ x\alpha z \text{ and } y\alpha z \text{ entails } x\alpha y \end{cases}$

Then there exists a unique class v and a unique function f which satisfies the following properties.

1. For any x in u, fx is an element of v;
2. For any h in v, there exists at least one element x of u such that $fx = h$;
3. For any x and y in u, one has $fx = fy$ if and only if x is in the relation α with y, that is $x\alpha y$.

It follows that v and f are determinate functions of u and α; the proposition $x\alpha y$ between the pairs x and y, infinite in number, of elements in u, is expressed by means of the unique identity $fx = fy$; for the infinite number of pairs of elements of u related by α, one can substitute a unique element of the class v.[37]

37 This last remark shows the practical importance of definition by abstraction which yields for instance the rational numbers, the real numbers, the complex numbers, as simple entities whereas for Mr. Russell they are *classes, classes of classes, ordered pairs of classes of classes*, which are very complicated to handle.

Let us apply for instance definition by abstraction in order to obtain the geometrical entities called *directions* (or *points at infinity*). The class u will be the class of *straight lines*; α will express the relation "*is parallel to*". Since the conditions in (3) are verified[38], the function f exists: we will call it *direction*. In what follows the class v of *directions* also exists.

In a similar way one obtains, starting from the relation "is parallel to" for planes, the class of *directions of planes* (or *lines at infinity*); from the relation "it can be superposed over" for segments, one obtains the *lengths*; from "is equivalent to" the *areas* and *volumes*; from "is similar to" the *shape* of a geometrical figure. And starting from suitably chosen relations one obtains the *weights*, *masses*, *temperatures*, *quantities of heat*,..., *the sense of a succession of 2, 3, 4 points* etc.

The *proper* definition (from a logical point of view) of *vectors* that we should have given [this refers to the textbook from which this appendix is taken, PM] is the following.

Let us consider the class u containing all the ordered pairs of points and the relation α defined as follows:

$$(A, B)\alpha(C, D)$$

if and only if

(4) "the middle point between A and D" = "middle point between B and C".

The relation (α) verifies the conditions (3). Consequently, the function f and the class v exist. Let us write, by definition,

$$f(A, B) = B - A$$

and let us call "vector from A to B" the entity $f(A, B)$, or $B - A$; in this way one obtains the definition of the class of vectors.

In the text, in order not to wander off too much from the usual form of definitions, we have written

$$B - A = D - C$$

when the relation (4) holds. We have thus *defined* in this way the relation = between two vectors. However, as we have remarked, this is no longer permitted, once one admits Leibniz's universal notion of equality.

38 The line a is parallel to the line b in the following two cases: 1) if a coincides with b; 2) if a and b are on the same plane and have no point in common. This is the definition that we adopt here and not Euclid's definition.

For what concerned the forms $\mathbf{F}_1$ (p. 17), the class u consists of all the pairs made up by a point and a real number; the relation α is the following:

$$(x_1, A_1; x_2, A_2; \ldots; x_n, A_n)\alpha(y_1, B_1; y_2, B_2; \ldots; y_m, B_m)$$

if and only if, for any arbitrary chosen point O, we have

$$\sum_1^n x_i(A_i - O) = \sum_1^m y_i(B_i - O)$$

The conditions expressed in (3) are verified etc. One can proceed in the same way for the forms $\mathbf{F}_2$, $\mathbf{F}_3$ (appendix, p. 181).

Uwe Meixner

Intelligible Worlds

Abstract: The paradigmatic mereological relation is the relation of spatial part. Already much less paradigmatic is the relation of temporal part. The realm of abstract entities seems to be the ontological region where the notion of part and whole has no application at all. In what follows, I will contend that this is not true. There are part-whole relationships between abstract entities, and indeed relationships that are systematic to the point of constituting mereologically structured universes of abstract entities, "intelligible worlds", as I will call them (in translation of the Latin "mundi intelligibiles"). The part-whole relations between abstract entities differ significantly from those between spatial, or temporal, or spatio-temporal entities. However, there are also significant analogies between abstract and concrete part-whole relations.

1 Preliminaries

The paradigmatic mereological relation is the relation of spatial part. Already much less paradigmatic is the relation of temporal part. The realm of abstract entities seems to be the ontological region where the notion of part and whole has no application at all. In what follows, I will contend that this is not true. There are part-whole relationships between abstract entities, and indeed relationships that are systematic to the point of constituting mereologically structured universes of abstract entities, "intelligible worlds", as I will call them (in translation of the Latin "mundi intelligibiles"). The part-whole relations between abstract entities differ significantly from those between spatial, or temporal, or spatio-temporal entities. However, there are also significant analogies between abstract and concrete part-whole relations, as we shall see.

The basic mereological language is a language of first-order predicate logic in which "(xPy)" and "$(x = y)$" (and all the variants of these two expressions that can be produced by employing *all manners* of replacing "x" and "y" in them by "x", "y", "z", "u", "v", "w", "x'", "y'", etc.) are the only *basic predicates*. The basic logical constants are $\neg$ (negation), $\rightarrow$ (material implication), $\forall$ (the all-quantifier) and ι (the operator of definite description). As is well known, this basis is sufficient for defining all truth-functional connectives, and in the first place $\wedge$, $\vee$, and $\leftrightarrow$, in other words: conjunction, non-exclusive disjunction,

Uwe Meixner: University of Augsburg

DOI: 10.1515/9783110529494-018

and material equivalence. In order to save brackets, it is stipulated that binding strength decreases from left to right in the following series: $\neg$, $\wedge$, $\vee$, $\rightarrow$, $\leftrightarrow$. And note that the embracing brackets in the *basic predicates* (and in the sentences formed from them by saturation with terms) will be omitted, unless the predicate – or a one-place predicate resulting from it by substitution of a term for a variable – constitutes the range of a quantifier, *or* the range of the operator of definite description (or of another term-forming operator), *or* the range of the negation-operator ($\neg$). Further bracket-saving measures, here implemented, are the following: Outer brackets – that is, such as occur if the expression enclosed by them is not within another expression – will always be omitted. Brackets within $\wedge$-chains and $\vee$-chains will always be omitted. As far as brackets are concerned, the defined predicate $\neq (x \neq y := \neg(x = y))$ is treated just like the basic predicate $=$.

The indicated basis also suffices to define all *at-most-N* quantifiers and all *at-least-N* quantifiers, and therefore also all *precisely-N* quantifiers (where N stands for any Arabic numeral designating a natural number). The most prominent *at-least-N* quantifier is the *at-least*-1 quantifier, or in other words, $\exists$, which is defined as follows: $\exists x A[x] := \neg\forall\neg A[x]$. The most prominent *precisely-N* quantifier is the *precisely*-1 quantifier, $\exists^{=1}$, which is defined as follows: $\exists^{=1} x A[x] := \exists x(A[x] \wedge \forall y(A[y] \rightarrow y = x))$.[1]

The logic employed is classical first-order logic with identity and definite descriptions. I will not bother to write down this logic, since it is well known. What deserves some attention, however, is the treatment here accorded to definite descriptions. The two relevant axiom-schemata are these: $\exists^{=1} x A[x] \rightarrow A[\iota x A[x]]$ and $\neg\exists^{=1} x A[x] \rightarrow \iota x A[x] = \iota y(y \neq y)$. Thus, a definite description $\iota x A[x]$ whose condition of normalcy $\exists^{=1} x A[x]$ is not fulfilled designates the same object as is designated by "$\iota y(y \neq y)$"; this object is some arbitrarily chosen object in the universe of discourse.

To the extent deductions and proofs are presented in what follows, these deductions and proofs are going to be informal (for the sake of readability). But, of course, they can be transposed into the strict or formal mode – if one is ready to undergo the trouble.

1 The variables "x" and "y" are used in this definition in a merely representative fashion. Other contexts will require the use of other variables. It does not matter which variables are employed as long as syntactic well-formedness and the structure required by the definitions is preserved. These observations apply to all definitions and also to all axioms and theorems that follow.

2 A mereology for abstract entities

First of all, here are two definitions of mereological predicates, and one definition of a mereological operator (all three defined expressions will be needed right away):

D1: $xP^*y := xPy \wedge \neg\forall u(xPu)$[2]
D2: $EL(z) := \forall x(xP^*z \rightarrow x = z)$
D3: $\sigma yA[y] := \iota u(\forall y(A[y] \rightarrow yPu) \wedge \forall x(\forall y(A[y] \rightarrow yPx) \rightarrow uPx))$

D1 defines what is meant by "*x is a non-trivial part of y*"; it is this: *x is a part of y without being a part of everything* (in the universe of discourse). **D2** defines what is meant by "*z is an elementary whole*"; it is this: *every non-trivial part of z is z*. Note that **D2** is not quite the definition of "$AT(z)$" (or: "*z is an atom*"); for the definition of this latter predicate is this:

D4: $AT(z) := \forall x(xPz \rightarrow x = z)$

In other words, an atom is something *that has no proper parts* (since $\neg\exists x(xPz \wedge \neg(x = z))$ is logically equivalent to $\forall x(xPz \rightarrow x = z)$. It is trivially provable that every atom is an elementary whole; the converse, however, is not provable.

D3, finally, defines what is meant by "*the sum of all y such that* $A[y]$"; it is this: *the mereologically smallest entity (in the universe of discourse) that comprises all entities (in the universe of discourse) that satisfy* $A[y]$. The principles **A4** and **A3** below guarantee for every predicate $A[y]$ (expressible in the language) that the condition of unique fulfilment is satisfied for the following predicate corresponding to $A[y]$: $\forall y(A[y] \rightarrow yPu) \wedge \forall x(\forall y(A[y] \rightarrow yPx) \rightarrow uPx)$. Thus, $\sigma yA[y]$ always refers to what, judging by its meaning (or sense), it is supposed to refer to.

Consider, then, the following series of axioms and axiom-schemata:

A1: $\forall x\forall y\forall z(xPy \wedge yPz \rightarrow xPz)$
A2: $\forall x(xPx)$
A3: $\forall x\forall y(xPy \wedge yPx \rightarrow x = y)$
A4: $\exists u(\forall y(A[y] \rightarrow yPu) \wedge \forall x(\forall y(A[y] \rightarrow yPx) \rightarrow uPx))$
A5: $\forall x\forall y(\forall z(EL(z) \wedge zPx \rightarrow zPy) \rightarrow xPy)$
A6: $\forall x(xP^*\sigma uA[u] \rightarrow \exists z(zP^*x \wedge \exists y(A[y] \wedge zP^*y)))$

2 Regarding embracing brackets, "(xP^*y)" acts just like "(xPy)".

The *natural* interpretation of this axiomatic theory is that it (truthfully) describes some *intelligible world* (in the sense introduced in section 1).[3] In fact, there are several candidates for what it *naturally describes*, as we shall see. But what is it that makes it appropriate to say that this mereology, **A1**–**A6**, is *naturally about* abstract entities? It is simply the fact that it is *not natural* to view it as a mereology for concrete entities. There are some features of it which make interpreting it as a mereology for concrete entities *unnatural* – indeed, which make such an interpretation *unfeasible* for paradigmatic concrete totalities, like real space and real time. This is already the case if one takes the mereology as it is, but it is most dramatically apparent if one adds existence assumptions that lift **A1**–**A6** above the level of trivial satisfiability.

Consider $\sigma u(u \neq u)$, in other words: $\iota u(\forall y(y \neq y \rightarrow yPu) \wedge \forall x(\forall y(y \neq y \rightarrow yPx) \rightarrow uPx))$. On the basis of **A4** and **A3**, it is easy to prove

T1: $\forall x(\sigma u(u \neq u)Px)$

and its corollary

T2: $\exists y \forall x(yPx)$

Obviously, it is not a natural mereological feature of concrete entities that there is an entity among them which is a part of all of them. If we look at *real* space, there is no spatial whole which is a spatial part of every spatial whole, and if we look at *real* time, there is no temporal whole which is a temporal part of every temporal whole. Thus **T2** (and therefore the conjunction of the principles of which **T2** is a logical consequence) is not true of spatial wholes, and not true of temporal wholes. In fact, even if space-points were counted as spatial wholes and there were only two space-points, there would be no spatial whole that is a part of every spatial whole; and even if time-points were counted as temporal wholes and there were only two time-points, there would be no temporal whole that is a part of every temporal whole.

Consider next *elementary wholes*, as defined by **D2**. If (using **D1**) we unpack the *definiens* of $EL(z)$ – $\forall x(xP^*z \rightarrow x = z)$ – and bring the result into a different but logically equivalent form, we obtain:

T3: $\forall z(EL(z) \leftrightarrow \forall x(xPz \wedge x \neq z \rightarrow \forall u(xPu)))$

3 What is (truthfully) described by a theory is called a "model" for it. A model for a theory can be artificially concocted, made up by applying ad hoc procedures and constructions; it can be specially sought out – or it can be simply *natural*.

T3 (a consequence of mere logic and definitions) says that the elementary wholes are precisely the entities all of whose proper parts are parts of every entity. This entails that *each elementary whole that is different from* $\sigma u(u \neq u)$ *has* $\sigma u(u \neq u)$ *as its one and only proper part*. *How* does this follow? Consider that it is precisely what is stated by **T6** below. But, first of all, we have **T4**:

T4: $\forall x'(\forall u'(x'Pu') \leftrightarrow x' = \sigma u(u \neq u))$

Proof. (I) Suppose $x' = \sigma u(u \neq u)$; hence by **T1**: $\forall u'(x'Pu')$. Suppose $\forall u'(x'Pu')$; by **T1**: $\forall x(\sigma u(u \neq u)Px)$; hence $x'P\sigma u(u \neq u) \wedge \sigma u(u \neq u)Px'$; hence by **A3**: $x' = \sigma u(u \neq u)$. qed

From **T3** and **T4** we get:

T5: $\forall z(EL(z) \leftrightarrow \forall x(xPz \wedge x \neq z \rightarrow x = \sigma u(u \neq u)))$

And therefore:

T6: $\forall z(EL(z) \wedge z \neq \sigma u(u \neq u) \rightarrow \sigma u(u \neq u)Pz \wedge \sigma u(u \neq u) \neq z \wedge \forall x(xPz \wedge x \neq z \rightarrow x = \sigma u(u \neq u)))$

Proof. Suppose $EL(z) \wedge z \neq \sigma u(u \neq u)$; hence according to **T1** (and the symmetry of non-identity): (i) $\sigma u(u \neq u)Pz \wedge \sigma u(u \neq u) \neq z$; and according to **T5**: (ii) $\forall x(xPz \wedge x \neq z \rightarrow x = \sigma u(u \neq u))$. qed

Now, evidently, neither spatial nor temporal wholes are entities that have exactly one proper part. Perhaps some of them have no proper parts, but certainly none of them have exactly one proper part. In fact, it is one of the most widespread mereological intuitions that if *any* entity y has a proper part x – and certainly there are such entities – that then it must also have at least one *other* proper part, namely, the *complement* of x relative to y; moreover, the *complementing* proper part of y is intuited to have no part in common with the complemented proper part of y. As convincing as this may sound (or rather *look*: one sees it "in the mind's eye"), it is nonetheless only true of *concrete* entities and *concrete* part-whole relations: For some *intelligible worlds*, not only **T6** is true but also $\exists z(EL(z) \wedge z \neq \sigma u(u \neq u))$ (as we shall see); the logical consequence of this is that, for such worlds, $\exists z \exists^{=1} x(xPz \wedge x \neq z)$ is also true – squarely contradicting the widespread mereological intuition. Moreover, if one follows **A1–A6**, then there simply are no complements as intended by the above-mentioned widespread intuition; because everything (in the universe of discourse) has a part in common with everything, due to **T1**.

And there is yet more food for wonder here. For some intelligible worlds, $\exists^{\geq 2} z(EL(z) \wedge z \neq \sigma u(u \neq u))$ is true (as we shall see); it follows on the basis of **T6** that there are *two* elementary wholes, *both* different from $\sigma u(u \neq u)$, which both have $\sigma u(u \neq u)$ as their sole proper part. How can this be? What distinguishes the two if they are identical with respect to proper parts? That there is something that distinguishes them is inconceivable for *concrete* entities; but for abstract entities it is quite a different matter (as we shall see).

Finally, if one hears of *atoms*, the immediate association is that there are *many* of them and that other entities – in fact, all other entities of a given kind – are *composed* of them, in such a manner that the sets of atoms that go into composing those other entities are different if the entities themselves are different. This is the intuitive view of atoms, which treats atoms as *concrete* entities. But on the basis of the above principles it turns out that there is only *one atom*, $\sigma u(u \neq u)$:

T7: $AT(\sigma u(u \neq u)) \wedge \forall z(AT(z) \rightarrow z = \sigma u(u \neq u))$

Proof. (I) Suppose $xP\sigma u(u \neq u)$; by **T1**: $\sigma u(u \neq u)Px$; hence by **A3**: $x = \sigma u(u \neq u)$. Therefore: $\forall x(xP\sigma u(u \neq u) \rightarrow x = \sigma u(u \neq u))$; hence by **D4**: $AT(\sigma u(u \neq u))$. (II) Suppose $AT(z)$; by **T1**: $\sigma u(u \neq u)Pz$; hence by supposition and **D4**: $\sigma u(u \neq u) = z$, hence $z = \sigma u(u \neq u)$. Therefore: $\forall z(AT(z) \rightarrow z = \sigma u(u \neq u))$. qed

Since there is only one atom (in the universe of discourse), nothing (in the universe of discourse) can be composed of atoms (*plural*). And if one allowed (departing from common usage, but not unacceptably) that something may also be *composed of just one atom*, then it is – according to **A1–A6**, and assuming $\exists^{\geq 2} z(EL(z) \wedge z \neq \sigma u(u \neq u))$ – still not true that different entities which are *composed of one atom* are each composed of a different atom: the various elementary wholes that differ from $\sigma u(u \neq u)$ are all composed of one atom, but it is always the same atom, $\sigma u(u \neq u)$, as we have already seen (consider the consequences of T6).

In **A1–A6**, the role of *atoms* is transferred to the *elementary wholes*. Not for the predicate $AT(z)$, but for the predicate $EL(z)$, it is provable.

T8: $\forall x(x = \sigma z(EL(z) \wedge zPx))$

Proof. (I) It is an easy consequence of **A4**, **A3**, and **D3**: $\forall u(EL(u) \wedge uPx \rightarrow uP\sigma z(EL(z) \wedge zPx))$; hence by **A5**: $xP\sigma z(EL(z) \wedge zPx)$. (II) Suppose $EL(u) \wedge uP\sigma z(EL(z) \wedge zPx))$; if $\forall x'(uPx')$, then uPx; if, on the other hand, $\neg\forall x'(uPx')$, then $uP^*\sigma z(EL(z) \wedge zPx))$ according to **D1**, and consequently by **A6**: $\exists z'(z'P^*u \wedge \exists y(EL(y) \wedge yPx \wedge z'P^*y))$; hence by logical transformations and by making use of the assumption $EL(u)$: $\exists z'\exists y(EL(u) \wedge EL(y) \wedge z'P^*u \wedge z'P^*y \wedge yPx)$; hence by **D2**: $\exists z'\exists y(EL(u) \wedge EL(y) \wedge z' = u \wedge z' = y \wedge yPx)$; hence uPx. It has now been

proven: $\forall u(EL(u) \wedge uP\sigma z(EL(z) \wedge zPx) \rightarrow uPx)$; hence by **A5**: $\sigma z(EL(z) \wedge zPx)Px$. By combining (I) and (II), it follows on the basis of **A3**: $x = \sigma z(EL(z) \wedge zPx)$. qed

T9: $\forall x \forall y(x \neq y \rightarrow \exists z(EL(z) \wedge zPx \wedge \neg(zPy)) \vee \exists z(EL(z) \wedge zPy \wedge \neg(zPx)))$

Proof. Proof: Suppose $x \neq y$; hence by **A3**: $\neg(xPy) \vee \neg(yPx)$. If the first alternative of this disjunction is true, then by **A5**: $\exists z(EL(z) \wedge zPx \wedge \neg(zPy))$; if the second alternative is true, then again by **A5**: $\exists z(EL(z) \wedge zPy \wedge \neg(zPx))$; hence in either case: $\exists z(EL(z) \wedge zPx \wedge \neg(zPy)) \vee \exists z(EL(z) \wedge zPy \wedge \neg(zPx))$. qed

Thus, every entity (in the universe of discourse) is the sum of its *elementary parts* (i.e., the sum of the elementary wholes that are parts of it), and if entities (in the universe of discourse) *differ*, then they differ with respect to at least one elementary part.

3 Complement, foundation, and top

If x is the sum of all elementary wholes that are parts of x, what is the sum of all elementary wholes that are not parts of x? – This latter sum is the complement of x:

D5: $com(x) := \sigma z(EL(z) \wedge \neg zPx)$

We have so far been looking at the *foundations* of intelligible worlds structurally defined by **A1–A6**; we now take a look at their *tops*. The tops are opposite to the foundations, or in other words: the tops are the complements of the foundations (and vice versa). To put it in an exact manner: the entities in a given top (of an intelligible world structurally defined by **A1–A6**), that is, *the comprehensive wholes* (among them $\sigma u(u = u)$), are precisely the complements of the entities in the foundation: they are the complements of the elementary wholes (among these $\sigma u(u \neq u)$).

The following definitions are the counterparts of **D1**, **D2**, and **D4**:

cD1: $xP^o y := xPy \wedge \neg \forall u(uPy)$
cD2: $CO(z) := \forall x(zP^o x \rightarrow x = z)$
cD4: $TO(z) := \forall x(zPx \rightarrow x = z)$

cD1 defines what is means for x to be a *distinguished part* of y: x is a part of y without everything (in the universe of discourse) being a part of y; **cD2** defines

what it means for z to be a *comprehensive whole*: every entity (in the universe of discourse) of which z is a distinguished part is identical to z; **cD4** defines what it means for z to be a *totality*: every entity (in the universe of discourse) of which z is a part is identical to z. The following theorems, then, are the counterparts of the theorems **T1–T9**:

cT1: $\forall x(xP\sigma u(u = u))$
cT2: $\exists y\forall x(xPy)$
cT3: $\forall z(CO(z) \leftrightarrow \forall x(zPx \wedge x \neq z \rightarrow \forall u(uPx)))$
cT4: $\forall x'(\forall u'(u'Px') \leftrightarrow x' = \sigma u(u = u))$
cT5: $\forall z(CO(z) \leftrightarrow \forall x(zPx \wedge x \neq z \rightarrow x = \sigma u(u = u)))$
cT6: $\forall z(CO(z) \wedge z \neq \sigma u(u = u) \rightarrow zP\sigma u(u = u) \wedge \sigma u(u = u) \neq z \wedge \forall x(zPx \wedge x \neq z \rightarrow x = \sigma u(u = u)))$
cT7: $TO(\sigma u(u = u)) \wedge \forall z(TO(z) \rightarrow z = \sigma u(u = u))$
cT8: $\forall x(x = \sigma z\forall y(CO(y) \wedge xPy \rightarrow zPy))$
cT9: $\forall x\forall y(x \neq y \rightarrow \exists z(CO(z) \wedge xPz \wedge \neg(yPz)) \vee \exists z(CO(z) \wedge yPz \wedge \neg(xPz)))$

The proofs of **cT1–cT9** (which I shall not present here) are somewhat harder to achieve than the proofs of **T1–T9**, since the principles **A1–A6** have an orientation towards the *foundations* of the intelligible worlds structurally defined by them, not towards their *tops*. In proving **cT1–cT9**, it is helpful to avail oneself of the following six theorems, which, taken together, establish a match between *tops* and *foundations*:

T10: $\forall x(EL(x) \wedge \neg\forall u(xPu) \rightarrow (xP\sigma z(EL(z) \wedge B[z]) \leftrightarrow B[x]))$

Proof. *Assume* $EL(x) \wedge \neg\forall u(xPu)$. (I) Suppose $B[x]$; hence by *the assumption*, **A4**, **A3**, **D3**: $xP\sigma z(EL(z) \wedge B[z])$. (II) Suppose $xP\sigma z(EL(z) \wedge B[z])$; hence by *the assumption* and **D1**: $xP^*\sigma z(EL(z) \wedge B[z])$; hence by **A6**: $\exists z'(z'P^*x \wedge \exists y(EL(y) \wedge B[y] \wedge z'P^*y))$; hence by logical transformations and *the assumption*: $\exists z'\exists y(EL(x) \wedge EL(y) \wedge z'P^*x \wedge z'P^*y \wedge B[y])$; hence by **D2**: $\exists z'\exists y(EL(x) \wedge EL(y) \wedge z' = x \wedge z' = y \wedge B[y])$; hence $B[x]$. qed

T11: $\forall x(com(com(x)) = x)$

Proof. (I) Suppose $EL(u) \wedge uPcom(com(x))$. If $\forall u'(uPu')$, then uPx. If $\neg\forall u'(uPu')$, then according to **T10**: $uP\sigma z(EL(z) \wedge \neg(zPcom(x))) \leftrightarrow \neg(uPcom(x))$, and therefore because of $uPcom(com(x))$ and **D5**: $\neg(uPcom(x))$; *and* then once more according to **T10**: $uP\sigma z'(EL(z') \wedge \neg(z'Px)) \leftrightarrow \neg(uPx)$, and therefore because of $\neg(uPcom(x))$ and **D5**: uPx. (II) Suppose $EL(u) \wedge uPx$. If $\forall u'(uPu')$, then $uPcom(com(x))$. If

$\neg\forall u^{\prime}(uPu^{\prime})$, then according to **T10** (as we have just seen): $uPcom(com(x)) \leftrightarrow \neg(uPcom(x)) \leftrightarrow uPx$, and therefore because of uPx: $uPcom(com(x))$. On the basis of (I) and **A5**, we have: $com(com(x))Px$; on the basis of (II) and **A5**, we have: $xPcom(com(x))$; on the basis of **A3**, we therefore obtain: $com(com(x)) = x$. qed

T12: $\forall x\forall y(xPy \leftrightarrow com(y)Pcom(x))$

Proof. (I) *Assume* xPy; suppose $EL(z^{\prime}) \wedge z^{\prime}Pcom(y)$; if $\forall u(z^{\prime}Pu)$, then $z^{\prime}Pcom(x)$; if $\neg\forall u(z^{\prime}Pu)$, then by **T10** and **D5** from $z^{\prime}Pcom(y)$: $\neg(z^{\prime}Py)$; hence by **A1** and *the assumption*: $\neg(z^{\prime}Px)$, hence by **A4**, **A3**, **D3**: $z^{\prime}P\sigma z(EL(z) \wedge \neg(z^{\prime}Px))$, hence by **D5**: $z^{\prime}Pcom(x)$. We have now established: $\forall z^{\prime}(EL(z^{\prime}) \wedge z^{\prime}Pcom(y) \rightarrow z^{\prime}Pcom(x))$; hence by A5: $com(y)Pcom(x)$. (II) Assume $com(y)Pcom(x)$; hence on the basis of what has already been established in (I) [the left-to-right part of **T12**]: $com(com(x))Pcom(com(y))$; hence on the basis of **T11**: xPy. qed

T13: $\forall z(CO(z) \leftrightarrow EL(com(z)), \forall z(CO(com(z)) \leftrightarrow EL(z)))$

Proof. (I) *Assume* $CO(z)$, hence by **cD2** and **cD1**: $\forall x(zPx \wedge \neg\forall u^{\prime}(u^{\prime}Px) \rightarrow x = z)$. Suppose $x^{\prime}Pcom(z) \wedge \neg\forall u^{\prime}(x^{\prime}Pu^{\prime})$; hence by **T12** and **T11**: $zPcom(x^{\prime}) \wedge \neg\forall u^{\prime}(com(u^{\prime})Pcom(x^{\prime}))$; hence $\neg\forall u^{\prime}(u^{\prime}Pcom(x^{\prime}))$ [for if $\forall u^{\prime}(u^{\prime}Pcom(x^{\prime}))$ were true, then certainly also $\forall u^{\prime}(com(u^{\prime})Pcom(x^{\prime}))$ would be true]. Therefore, on the basis of *the assumption*, we have: $com(x^{\prime}) = z$, hence: $com(com(x^{\prime})) = com(z)$, hence by **T11**: $x^{\prime} = com(z)$. We have now seen: $\forall x^{\prime}(x^{\prime}Pcom(z) \wedge \neg\forall u^{\prime}(x^{\prime}Pu^{\prime}) \rightarrow x^{\prime} = com(z))$, hence by **D1** and **D2**: $EL(com(z))$. (II) *Assume* $EL(com(z))$, hence by **D2** and **D1**: $\forall x(xPcom(z) \wedge \neg\forall u^{\prime}(xPu^{\prime}) \rightarrow x = com(z))$. Suppose $zPx^{\prime} \wedge \neg\forall u^{\prime}(u^{\prime}Px^{\prime})$; hence by **T12**: $com(x^{\prime})Pcom(z) \wedge \neg\forall u^{\prime}(com(x^{\prime})Pcom(u^{\prime}))$; hence $\neg\forall u^{\prime}(com(x^{\prime})Pu^{\prime})$ [for if $\forall u^{\prime}(com(x^{\prime})Pu^{\prime})$ were true, then certainly also $\forall u^{\prime}(com(x^{\prime})Pcom(u^{\prime}))$ would be true]. Therefore, on the basis of *the assumption*, we have: $com(x^{\prime}) = com(z)$, hence: $com(com(x^{\prime})) = com(com(z))$, hence by **T11**: $x^{\prime} = z$. We have now seen: $\forall x^{\prime}(zPx^{\prime} \wedge \neg\forall u^{\prime}(u^{\prime}Px^{\prime}) \rightarrow x^{\prime} = z)$, hence by **cD1** and **cD2**: $CO(z)$. The second part of **T13** is an easy corollary of the first part, given **T11**. qed

T14: $\sigma u(u = u) = com(\sigma u(u \neq u))$

Proof. (I) Because of **cT1**: $com(\sigma u(u \neq u))P\sigma u(u = u)$. (II) *Assume* $EL(z^{\prime}) \wedge z^{\prime}P\sigma u(u = u)$; if $\forall u^{\prime}(z^{\prime}Pu^{\prime})$, then $z^{\prime}Pcom(\sigma u(u \neq u))$; if $\neg\forall u^{\prime}(z^{\prime}Pu^{\prime})$, then $\neg(z^{\prime}P\sigma u(u \neq u))$,[4] and therefore: $z^{\prime}P\sigma z(EL(z) \wedge \neg(zP\sigma u(u \neq u)))$, on the

4 If $z^{\prime}P\sigma u(u \neq u)$, then $z^{\prime} = \sigma u(u \neq u)$ (because of **T7** and **D4**), and consequently $\forall u^{\prime}(z^{\prime}Pu^{\prime})$ because of **T1**.

basis of **A4**, **A3**, **D3**; hence $z^{I}Pcom(\sigma u(u \neq u))$ because of **D5**. We have now established: $\forall z^{I}(EL(z^{I}) \wedge z^{I}P\sigma u(u = u) \rightarrow z^{I}Pcom(\sigma u(u \neq u)))$; hence by **A5**: $\sigma u(u = u)Pcom(\sigma u(u \neq u))$. Given (I) and (II), **T14** follows by **A3**. qed

T15: $\forall x(CO(x) \leftrightarrow \exists y(EL(y) \wedge x = com(y))), \forall x(EL(x) \leftrightarrow \exists y(CO(y) \wedge x = com(y)))$

Proof. (I) Assume $CO(x)$; hence by **T13**: $EL(com(x))$; hence by **T11**: $EL(com(x)) \wedge x = com(com(x))$; hence $\exists y(EL(y) \wedge x = com(y))$. (II) Assume $\exists y(EL(y) \wedge x = com(y))$; by **T13**: $\exists y(EL(y) \wedge CO(com(y)) \wedge x = com(y))$; hence $CO(x)$. The proof of the second part of **T15** is entirely analogous. qed

4 Models for A1–A6

When we look at the contents of the theorems **cT1–cT9**, it turns out that part-whole-relations between certain *concrete* entities are *to some extent* as blatantly out of accord with what *those* theorems are implying as they are out of accord with what **T1–T9** are implying. For example, one will not find a spatial whole (that is, a part of real space) that differs from the spatial totality (that is, from real space) in such a manner that it is a proper part *only* of the spatial totality; at least this is true if one does not count space-points as spatial wholes.[5] And one will not find a temporal whole (that is, a part of real time) that differs from the temporal totality (real time) in such a manner that it is a proper part only of the temporal totality. It is true that **A1–A6** do not entail that there is a whole that differs from the totality in such a manner that it is a proper part only of the totality. But the mere extra assumption $\exists z(CO(z) \wedge z \neq \sigma u(u = u))$ ("There is at least one comprehensive whole that differs from the totality") will yield $\exists z(zP\sigma u(u = u) \wedge \sigma u(u = u) \neq z \wedge \forall x(zPx \wedge x \neq z \rightarrow x = \sigma u(u = u)))$ on the basis of **cT6**.

It is, however, not without good reason that I put an emphasis on the phrase "to some extent" in the first sentence of this section (section 4). There *are* concrete totalities (each unique in the relevant model) which are such that some of their proper parts are proper parts only of them (in the relevant model). Consider a group G, consisting of four people; let G be *the* totality. Clearly, G is a concrete, non-abstract entity, and so are all of its subgroups (whether or not the *members* of G – the four people themselves – are counted as subgroups of G, that is, as

5 If one does count space-points as spatial wholes, then one can say that real space *without* a certain (arbitrary) space-point is a spatial whole of the envisaged kind.

singleton subgroups of G). It is evident that each of the four *three-membered* subgroups of G differs from G in such a manner that it is a proper part (proper subgroup) only of G. Moreover, it is easily seen that, if the universe of discourse encompasses G and all of its subgroups (of people) and nothing else, then all the theorems in **cT1–cT9** turn out to be true – given that "xPy" and "$\sigma u(u = u)$" are understood in the straightforward sense that the stipulated universe of discourse suggests.[6]

Readers may wonder whether the mereological model for **cT1–cT9** that has G for its totality – in short: the G-model – satisfies not only **cT1–cT9** but also **T1–T9**, because it simply satisfies **A1–A6**. If that were true, then there would be a *concrete* and rather natural model for a mereology that – at first – looked as if it was naturally appropriate only for intelligible worlds. To decide the matter, one has to be clear on the question of which entities, precisely, are in the stipulated universe of discourse. It comprises at least G, the four three-membered subgroups of G, and the six two-membered subgroups of G. Does it comprise anything else? Since the stipulated universe of discourse comprises G and all subgroups of G and nothing else, further candidates for being in the universe of discourse can only be one-membered and zero-membered subgroups of G (consisting of members of G: certain people). But an empty subgroup of G – a group which would be a subgroup of every subgroup of G – is out of the question, and singleton subgroups of G – each to be identified with one of the four *members* of G – are *groups* only by courtesy. In the strict acceptation of the word "group", there is nothing else in the universe of discourse than the already mentioned eleven entities; in a liberal acceptation of "group", four singleton subgroups of G are in the universe of discourse *in addition* to the eleven entities already mentioned.

Let us adopt the liberal position. The effect of this is that **A1–A3**, **A5** and **A6** turn out to be true; but **A4**, as it stands, cannot be true for the G-model; only **A4′** is true for it: $\exists y A[y] \rightarrow \exists u(\forall y(A[y] \rightarrow yPu) \wedge \forall x(\forall y(A[y] \rightarrow yPx) \rightarrow uPx))$.[7] Therefore,

---

6 G is the group which consists of Andrew, Anna, Nina, and Vladimir. The group which consists of Anna and Nina is a proper part of G, and so is the group which consists of Anna and Andrew. The (intended mereological) sum of these two proper parts of G is the group which consists of Anna, Nina, and Andrew, which group, too, is a proper part of G. The sum of all (self-identical) entities in the universe of discourse is certainly G. (According to the strict view, the number of those entities is 11; according to the liberal view, their number is 15.)

7 Thus, in the axiom-system whose models are the models that are *just like* the G-model, only **A4** needs to be replaced (by **A4′**), whereas **A1–A3**, **A5** and **A6** can be retained. However, certain simplifications are recommendable: In **A5**, "$EL(z)$" should be replaced by "$AT(z)$", and in **A6**, "P^*" should be replaced by "P". These simplifications are possible in view of **D1**, **D2**, and **D4**, and the fact that for the models that are *just like* the G-model (they contain at least one proper

the G-model is after all not a *concrete* natural model for **A1–A6**. But there certainly are *abstract* natural models for **A1–A6**. An entirely commonplace natural abstract model for **A1–A6** is obtained by stipulating that the universe of discourse is to contain all the subsets of a certain set S, and nothing else (it does not matter which set S is, it may even be the empty set), and by interpreting "xPy" as "x is a subset of y". Then the *elementary wholes* (the entities that satisfy "$EL(x)$") turn out to be the singleton subsets of S *plus* the empty set; and the *comprehensive wholes* (the entities that satisfy "$CO(x)$") turn out to be S *plus* the subsets of S that differ from S only by lacking precisely one element of S ("element" being taken in the set-theoretical sense).

The abstract natural models for **A1–A6** become more interesting if one adds an axiom-schema of infinity to **A1–A6**, for example in the following way:

A7: $\exists^{\geq 1} z(EL(z) \wedge \neg AT(z)) \wedge (\exists^{\geq N} z(EL(z) \wedge \neg AT(z)) \rightarrow \exists^{\geq N+1} z(EL(z) \wedge \neg AT(z)))$

Instead of $\exists^{\geq 1} z(EL(z) \wedge \neg AT(z)) \wedge (\exists^{\geq N} z(EL(z) \wedge \neg AT(z)) \rightarrow \exists^{\geq N+1} z(EL(z) \wedge \neg AT(z)))$, one can just as well choose $\exists^{\geq 1} z(CO(z) \wedge \neg TO(z)) \wedge (\exists^{\geq N} z(CO(z) \wedge \neg TO(z)) \rightarrow \exists^{\geq N+1} z(CO(z) \wedge \neg TO(z)))$ as axiom-schema of infinity. For on the basis of **A1–A6**, the former schema and the latter are deductively equivalent: whichever of the two schemata one chooses as the one which is to be axiomatic, one will be able to obtain the other one as a theorem.

Let the universe of discourse comprise, then, all the subsets of the set of natural numbers and nothing else, with "xPy" being interpreted as "x is a subset of y". This stipulation, obviously, provides us with an abstract natural model for **A1–A6** *plus* **A7**. The most interesting natural abstract models for **A1–A6** *plus* **A7** are, however, the following two: (I) Let the universe of discourse comprise *all states of affairs* and nothing else, with "xPy" being interpreted as "x is intensionally contained in y" (for example, the state of affairs that Peter is born earlier than John is intensionally contained in the state of affairs that John is born later than Peter, and the state of affairs that Peter has a date of birth is intensionally contained in the state of affairs that Peter is born earlier than John. (II) Let the universe of discourse comprise *all properties of individuals* and nothing else, with "xPy" being interpreted as "x is intensionally contained in y" (for example, the property of having a colour is intensionally contained in the property of being red, and the property of being extended is intensionally contained in the property of having a colour). If one accepts the world of states

– that is, *at least two-membered* – group and no empty group), $\forall x \neg \forall u(xPu)$ is always true; this fact makes xP^*y equivalent to xPy, and $EL(z)$ equivalent to $AT(z)$.

of affairs and the world of properties of individuals (both are *mundi intelligibiles*) as universes (of discourse) that conform to the descriptions provided by **A1–A6** plus **A7**, then this presupposes that one has made, in both cases, two momentous decisions *in addition* to the, doubtless, momentous decision to accept states of affairs and properties of individuals in huge numbers: one has decided to accept that entities which intensionally contain each other (be they states of affairs or properties of individuals) are identical to each other, that is, one has opted for a "coarse-grained" individuation of states affairs and properties of individuals (otherwise **A3** would be violated); and one has decided to accept that, with each state of affairs and each property of individuals, also its complement – or: its *negation*, as one says if talk is about states of affairs or properties – is a state of affairs, respectively, property of individuals. Each of these – in all – three decisions has been severely disapproved of by this or that philosopher. Yet, *if* one accepts abstract entities at all, and *if* one considers states of affairs and properties to be *abstract* entities, then – within the ontological framework defined by these two conditions (in fact, they point to yet further decisions) – all of the metaphysical decisions mentioned seem perfectly all right.

What **A1–A6** plus **A7** mean for states of affairs and for properties of individuals is explored in great detail (albeit in a somewhat different terminology) in my books *Axiomatic Formal Ontology* and *The Theory of Ontic Modalities*. Here, I would merely like to point out a few fascinating consequences which this formal mereological theory has for states of affairs and properties of individuals (taken to be abstract entities). Already in **A1–A6** the following theorem is provable:

T16: $\forall x(CO(x) \wedge \neg TO(x) \leftrightarrow \forall y(yPx \leftrightarrow \neg(com(y)Px)))$

T16 says that the comprehensive wholes which are not totalities – in other words (in view of **cT7**), the comprehensive wholes which are different from $\sigma u(u = u)$ – are precisely the *mereologically maximal-consistent wholes*, where a mereologically maximal-consistent whole is defined as an entity such that for each entity (in the universe of discourse) it is true that either that entity itself or its complement (but not both) is a part of it. Given **A7**, the number of comprehensive wholes which are not totalities – that is (by **T16**), the number of maximal-consistent wholes – is *infinite* (since there are precisely as many comprehensive wholes which are not totalities as there are elementary wholes which are not atoms, as can be proven in **A1–A6**: **T13**, second part, **T14**, and **T15**, first part, can be used as lemmas in the proof).

What are the maximal-consistent wholes if the entities in the universe of discourse are precisely the states of affairs? They are the *possible worlds*, *in abstracto* conceived of as maximal-consistent states of affairs (developing an

idea that can be gathered from Wittgenstein's *Tractatus*). And what are the maximal-consistent wholes if the entities in the universe of discourse are precisely the properties of individuals? In that case, they are the *notiones completae* of Leibniz, conceived of as maximal-consistent properties of individuals, each *notio completa* being the sum of all the properties a given individual has in a given possible world. The metaphysically profound question is whether there is an *essential* one-to-one match between individuals and *notiones completae* (qua maximal-consistent properties of individuals), or not. This question has two parts: (A) Does necessarily each *notio completa* have an individual as *its one and only exemplifier*, such that, necessarily, different *notiones* have different individuals as their sole exemplifiers, and such that necessarily there is for each individual a *notio completa* which has it as its sole exemplifier? (B) May a *notio completa* have a certain individual x as exemplifier *without* this being necessarily so? If question (A) is answered by "yes" and question (B) by "no", then there is indeed an essential one-to-one match between individuals and *notiones completae*, and one might as well identify the individuals (disregarding concreteness) with the *notiones completae*: the maximal-consistent properties of individuals. Among the interesting consequences of making this identification would be, for example, (i) the exemplification of a property by an individual – or in other words: the having of a property by an individual – turns into a single-category mereological relation: $xEXEMy := CO(x) \wedge \neg TO(x) \wedge yPx$; and (ii) the intuition that *an actual individual x could have had other properties than it really has* can only be accommodated by saying that what is really (literally) meant by this is the following: *a counterpart of x* (a certain maximal-consistent property) has (comprises) other properties than x, but is not actual.[8]

5 The geography of A1-to-A6 worlds

For each intelligible world W which conforms to (the descriptions provided by) **A1–A6** the following is true: the number of entities in W is $2^{c(EL\&\neg AT)}$, where $c(EL\&\neg AT)$ is the number of elementary wholes in W that are not atoms. $c(EL\&\neg AT)$ is taken from $0, 1, 2, 3, \ldots; \aleph_0$. Each intelligible **A1**-to-**A6** world with $1 \leq c(EL\&\neg AT)$ has two distinct *poles*: a *south pole*: $\sigma u(u \neq u)$, and a *north pole*: $\sigma u(u = u)$, with $\sigma u(u \neq u) \neq \sigma u(u = u)$. Each **A1**-to-**A6** world with $2 \leq c(EL\&\neg AT)$

8 For more on the application of "actual" to properties of individuals and states of affairs, see my books Meixner (1997) and Meixner (2006).

has at least one *latitude* between the two poles. If $3 \leq c(EL\&\neg AT)$, then the number of latitudes between the poles is ≥ 2 and the number of *northern latitudes* is equal to the number of *southern latitudes*. If $2 \leq c(EL\&\neg AT)$ and $c(EL\&\neg AT)$ is an even number, then there is an *equator*: a latitude which is neither a southern nor a northern latitude, but the border between the southern and the northern half of the world concerned. Each entity in an **A1**-to-**A6** world is either the south pole, or the north pole, or is in one of the latitudes of the intelligible world. No entity in a higher (more northern) latitude is ever part of an entity in a lower (more southern) latitude. The south pole is the entity (in the world concerned) that consists of *no* non-atomic elementary wholes (of the world concerned). In the first latitude *above* the south pole, there are the entities which consist of *precisely one* non-atomic elementary whole; in the second latitude above the south pole, there are the entities which consist of *precisely two* non-atomic elementary wholes; …; in the second latitude *below* the north pole, there are the entities which consist of *all but two* non-atomic elementary wholes; in the first latitude below the north pole, there are the entities which consist of *all but one* non-atomic elementary wholes. The north pole is the entity which consists of *all* non-atomic elementary wholes. The complement of the south pole is the north pole; the complement of an entity in the Nth latitude above the south pole is in the Nth latitude below the north pole; the complement of an entity in an equator is – *in the equator*.

Below, are the distribution schemata of entities in **A1**-to-**A6** worlds of the first seven cardinalities. Each summand in the sum-expressions stands for the number of entities to be found at the respective latitude or pole; the first summand (at the left) refers to the south pole, the last summand (at the right) to the north pole, the summands in between refer to the latitudes between the poles, one after the other; the central summand – if there is one – refers to the equator:

$2^0 = 1$
$2^1 = 1 + 1$
$2^2 = 1 + 2 + 1$
$2^3 = 1 + 3 + 3 + 1$
$2^4 = 1 + 4 + 6 + 4 + 1$
$2^5 = 1 + 5 + 10 + 10 + 5 + 1$
$2^6 = 1 + 6 + 15 + 20 + 15 + 6 + 1$
…

Consider again the natural model for **A1–A6** plus **A7** which has precisely the subsets of the set of natural numbers in the universe of discourse, with "xPy" being interpreted as "x is a subset of y". The world of this model has, besides the two poles (the south pole is the empty set, the north pole the set of natural

numbers), a denumerably infinite number of southern latitudes, each of them occupied by a denumerably infinite number of finite sets (first singletons, then pairs, then triples, then …); and it has a denumberably infinite number of northern latitudes, each of them occupied by a denumerably infinite number of denumerably infinite sets; and it has an equator, occupied by a superdenumerably infinite number of denumerably infinite sets.

6 Other intelligible worlds

The system **A1–A6** plus **A7** is certainly sufficient for determining that *natural* models of it are *abstract*, in other words, *intelligible worlds*. It is, however, not the case that every infinite intelligible world can serve as a model of **A1–A6** plus **A7**. Obviously, neither *the world of natural numbers* nor *the world of pure sets*[9] satisfies **A1–A6** (though there are countless sub-regions of the world of pure sets that satisfy **A1–A6** *and* **A7**). Just for the sake of curiosity: Which axiomatic system could serve as a mereology for the world of natural numbers (which world must be carefully distinguished from *the world of the sets of natural numbers*)? For obtaining such a mereology, the natural step is to interpret "xPy" as "$x \leq y$". This immediately yields the principles **A1–A3**, which, since the intended interpretation is now very different from the interpretation originally intended, are re-named into **B1–B3**:

B1: $\forall x \forall y \forall z(xPy \wedge yPz \rightarrow xPz)$
B2: $\forall x(xPx)$
B3: $\forall x \forall y(xPy \wedge yPx \rightarrow x = y)$

The linearity of the world of natural numbers is captured in a mereological way (given **B1** and **B3**) by the following principle (which principle makes **B2** superfluous: **B2** is straightforwardly deducible from it):

B4: $\forall x \forall y(xPy \vee yPx)$

The infinity and the discreteness of the world of natural numbers (given **B1**, **B3**, and **B4**) is captured in a mereological way by the following principle:

B5: $\forall x \exists z(xPz \wedge x \neq z \wedge \neg \exists z'(xPz' \wedge x \neq z' \wedge z'Pz \wedge z' \neq z))$

9 *Puresets* are the sets – conforming to a chosen axiomatic set theory – that would be still around if there were nothing else but sets.

On the basis of **B5** and **B4**, it is provable:

T′1: $\forall x\exists^{=1}z(xPz \wedge x \neq z \wedge \neg\exists z'(xPz' \wedge x \neq z' \wedge z'Pz \wedge z' \neq z))$

Proof. All that remains to be done in view of **B5** is to demonstrate uniqueness. *Assume*, therefore, for *reductio*: $xPz \wedge x \neq z \wedge \neg\exists z'(xPz' \wedge x \neq z' \wedge z'Pz \wedge z' \neq z) \wedge xPu \wedge x \neq u \wedge \neg\exists z'(xPz' \wedge x \neq z' \wedge z'Pu \wedge z' \neq u) \wedge u \neq z$. Because of **B4**: $zPu \vee uPz$. If zPu, then $xPz \wedge x \neq z \wedge zPu \wedge z \neq u$ – contradicting $\neg\exists z'(xPz' \wedge x \neq z' \wedge z'Pu \wedge z' \neq u)$. If, on the other hand, uPz, then $xPu \wedge x \neq u \wedge uPz \wedge u \neq z$ – contradicting $\neg\exists z'(xPz' \wedge x \neq z' \wedge z'Pz \wedge z' \neq z)$. qed

D′1: $succ(x) := \iota z(xPz \wedge x \neq z \wedge \neg\exists z'(xPz' \wedge x \neq z' \wedge z'Pz \wedge z' \neq z))$

D′1 defines the all-important successor-functor for natural numbers. The following *Peano-axiom* is a theorem of the present system:

T′2: $\forall x\forall y(succ(x) = succ(y) \rightarrow x = y)$

Proof. Assume $succ(x) = succ(y)$. By **T′1** and **D′1**: $xPsucc(x) \wedge x \neq succ(x) \wedge \neg\exists z'(xPz' \wedge x \neq z' \wedge z'Psucc(x) \wedge z' \neq succ(x))$ and $yPsucc(y) \wedge y \neq succ(y) \wedge \neg\exists z'(yPz' \wedge y \neq z' \wedge z'Psucc(y) \wedge z' \neq succ(y))$, hence by logical transformations: (i) $\forall z'(xPz' \wedge z'Psucc(x) \wedge z' \neq succ(x) \rightarrow x = z')$ and (ii) $\forall z'(yPz' \wedge z'Psucc(y) \wedge z' \neq succ(y) \rightarrow y = z')$. Now, by **B4**: $xPy \vee yPx$. In the first case, $xPy \wedge yPsucc(x)$ [since $yPsucc(y)$ and $succ(x) = succ(y)$] $\wedge$ $y \neq succ(x)$ [since $y \neq succ(y)$ and $succ(x) = succ(y)$], and therefore on the basis of (i): $x = y$. In the second case, $yPx \wedge xPsucc(y)$ [since $xPsucc(x)$ and $succ(x) = succ(y)$] $\wedge$ $x \neq succ(y)$ [since $x \neq succ(x)$ and $succ(x) = succ(y)$], and therefore on the basis of (ii): $y = x$, hence $x = y$. qed

Consider next the following two axiom-schemata (which are immediately evident in view of the intended interpretation):

B6a: $\exists zA[z] \rightarrow \exists u(A[u] \wedge \forall z(A[z] \rightarrow uPz))$
B6b: $\exists^{=N}zA[z] \rightarrow \exists u(A[u] \wedge \forall z(A[z] \rightarrow zPu))$
(where "*N*" stands for any Arabic numeral *except* "0")[10]

10 The mere use of Arabic numerals (as in $\exists^{=1}zA[z]$, $\exists^{=2}zA[z]$, $\exists^{=3}zA[z]$, ...) does not mean that one is using or presupposing arithmetic: $\exists^{=N}zA[z]$ is definable entirely without the use of arithmetic.

Using **B3**, it is easy to deduce the following theorems from **B6a** and **B6b**:

T′3a: $\exists z A[z] \to \exists^{=1} u(A[u] \wedge \forall z(A[z] \to uPz))$
T′3b: $\exists^{=N} z A[z] \to \exists^{=1} u(A[u] \wedge \forall z(A[z] \to zPu))$

And we have the following definitions:

D′2a: $\nu x A[x] := \iota u(A[u] \wedge \forall z(A[z] \to uPz))$
D′2b: $\sigma x A[x] := \iota u(A[u] \wedge \forall z(A[z] \to zPu))$

$\nu x A[x]$ is *the mereological nucleus* of the natural numbers that satisfy the predicate $A[u]$, in other words: $\nu x A[x]$ is the smallest natural number that satisfies $A[u]$; $\sigma x A[x]$ is *the mereological sum* of the natural numbers that satisfy the predicate $A[u]$, in other words: $\sigma x A[x]$ is the largest natural number that satisfies $A[u]$ (obviously, the mereological sum of natural numbers is not the arithmetical sum of them). An expression of the form $\nu x A[x]$ is not guaranteed to have, for just any predicate $A[x]$, a referent that conforms to its meaning; it is guaranteed to have such a referent only for predicates $A[x]$ for which $\exists z A[z]$ is true (see **T′3a**). In turn, an expression of the form $\sigma x A[x]$ is not guaranteed to have, for just any predicate $A[x]$, a referent that conforms to its meaning; it is guaranteed to have such a referent only for predicates $A[x]$ for which $\exists^{=N} z A[z]$ is true (see **T′3b**).

The following important theorems can now be proven, which show that the mereology of natural numbers is, after all, a mereology for *abstract entities* in a manner which is *to some extent analogous* to the way in which **A1–A6** plus **A7** is a mereology for abstract entities. According to these theorems, there is a *part of everything which*, at the same time, *is the one and only atom*; there is no natural concrete model for such a proposition.

T′4: $\forall z(\nu x(x = x)Pz)$

Proof. On the basis of **T′3a**, **D′2a**, and the (provable) logical truth $\exists x(x = x)$, we obtain (using the logic of definite descriptions): $\forall z(z = z \to \nu x(x = x)Pz)$; hence because of $\forall z(z = z)$: $\forall z(\nu x(x = x)Pz)$. qed

T′5: $\neg\exists z(zP\nu x(x = x) \wedge z \neq \nu x(x = x))$

Proof. If $zP\nu x(x = x)$, then it follows because of **T′4** and **B3**: $z = \nu x(x = x)$. qed

T′6: $\forall y(\neg\exists z(zPy \wedge z \neq y) \to y = \nu x(x = x))$

Proof. Assume $\neg\exists z(zPy \wedge z \neq y)$; by **T′4**: $\nu x(x = x)Py$; hence $\nu x(x = x) = y$, hence $y = \nu x(x = x)$. qed

Given the first definition in the following series of definitions,

D′3: $0 := \nu x(x = x)$, $1 := succ(0)$, $2 := succ(1)$, $3 := succ(2)$, etc.[11]

another Peano-axiom is easily provable:

T′7: $\neg\exists y(succ(y) = 0)$

Proof. Suppose $succ(y) = 0$; hence by **D′3**: $succ(y) = \nu x(x = x)$. By **T′1**, **D′3**: $yPsucc(y) \wedge y \neq succ(y)$. Hence $yP\nu x(x = x) \wedge y \neq \nu x(x = x)$ – contradicting **T′5**. qed

But what about the *central* Peano-axiom, the schema of complete induction? The schema of complete induction is directly assumed in the present system,

B7: $A[0] \wedge \forall x(A[x] \rightarrow A[succ(x)]) \rightarrow \forall xA[x]$,

since there appears to be no more perspicuous way than **B7** to describe the aspect of the world of natural numbers that **B7** is aiming at – except, perhaps, the *infinite* axiom $\forall x(x = 0 \vee x = 1 \vee x = 2 \vee \ldots \vee x = N \vee \ldots)$, taken to cover all and only expressions N that are definable in the way indicated in **D′3**. This axiom, however, is an infinitely long expression (requiring an infinitistic logic); it is, therefore, *non-standard*. With $\forall x(x = 0 \vee x = 1 \vee x = 2 \vee \ldots \vee x = N \vee \ldots)$ in place, **B7** is easily provable (employing infinitistic logic): Assume $A[0] \wedge \forall x(A[x] \rightarrow A[succ(x)])$; hence (using **D′3**): $A[0], A[1], A[2], \ldots, A[N], \ldots$; hence: $\forall x(x = 0 \rightarrow A[x])$, $\forall x(x = 1 \rightarrow A[x])$, $\forall x(x = 2 \rightarrow A[x])$, ..., $\forall x(x = N \rightarrow A[x])$, ...; hence: $\forall x(x = 0 \vee x = 1 \vee x = 2 \vee \ldots \vee x = N \vee \ldots \rightarrow A[x])$; hence because of $\forall x(x = 0 \vee x = 1 \vee x = 2 \vee \ldots \vee x = N \vee \ldots)$: $\forall xA[x]$.

11 Alternatively one could define: $0 := \nu x(x = x)$, $1 := \nu x(x \neq 0)$, $2 := \nu x(x \neq 0 \wedge x \neq 1)$, $3 := \nu x(x \neq 0 \wedge x \neq 1 \wedge x \neq 2)$, etc., and then prove: $1 = succ(0)$, $2 = succ(1)$, $3 = succ(2)$, etc. For example, "$1 = succ(0)$" is proven as follows: Since $\forall y(y \neq 0 \rightarrow \nu x(x \neq 0)Py)$ and $succ(0) \neq 0$, we have: $\nu x(x \neq 0)Psucc(0)$; and secondly we have: $0P\nu x(x \neq 0) \wedge 0 \neq \nu x(x \neq 0)$; and thirdly we have: $\neg\exists z'(0Pz' \wedge 0 \neq z' \wedge z'Psucc(0) \wedge z' \neq succ(0))$. Therefore: $\nu x(x \neq 0) = succ(0)$, hence: $1 = succ(0)$.

Carlo Nicolai

Necessary Truths and Supervaluations

1 Hierarchies of theories and evidences

Logical complexity is one of the most fascinating and deep facts stemming from the incompleteness phenomena, and it is also one of the main themes of Sergio Galvan's ongoing journey into logic and philosophy. Just to mention a well-known example, the complexity of the set of elementary *truths* of a first-order theory[1] containing a modicum of arithmetic will always exceed – in a formally precise sense – the complexity of the set of *theorems* of that theory.

The mismatch between truth and provability is one of the central research interests of Sergio Galvan, as it became clear already with his first work on Tarski (Galvan, 1973). The incompleteness theorems determine a hierarchy of 'natural' theories given by consistency strength or similar means of comparison. The consistency of Zermelo-Fraenkel set theory with choice ZFC can be proved for instance in ZFC plus the existence of the least wordly cardinal, which will then occupy a higher position than ZFC in the hierarchy. More generally, it is a consequence of Gödel's results that the consistency of a sufficiently rich T can only be proved in theories 'stronger' than T. Suitable set existence axioms, but also reflection principles and truth principles, have all been employed to properly extend T to theories that are capable of deriving its consistency or equivalent statements.

As Galvan (1992) lucidly points out, there is little epistemological interest in justifying the acceptance of T under assumptions stronger than T, at least if the kind of justification we are after is close to a fully-fledged *foundation*. Galvan's analysis of incompleteness, therefore, suggests to read-off, in the hierarchy of theories given by pure strength, a finer-grained hierarchy of *explanation*. To this hierarchy belong theories that are capable of formalizing and making explicit our commitment to theories lying lower down in the hierarchy. The theory PA+Con(PA) (cf. §2), obtained by adding a (intensionally correct) consistency statement to PA, will *not* belong to Galvan's hierarchy of explanation, although its consistency strength trivially exceeds the one of PA; the simple assumption of the consistency of PA does not represent in fact an explanation of our acceptance of PA, in Galvan's sense, but a mere stipulation. By contrast, the subsystem of second-order

1 Here by theory we always intend theory in classical logic.

Carlo Nicolai: University of Utrecht

DOI: 10.1515/9783110529494-019

arithmetic ACA will belong to the hierarchy of explanation as it can define a full truth predicate for PA – more specifically, a full truth class for PA: this suffices for formalizing in ACA the metatheoretic proof of the soundness of PA.

It is therefore natural to assume that one way to climb up Galvan's hierarchy of explanation, given a trustworthy starting point T, is to assume a theory of truth for it. In this way one may achieve a sort of 'explanatory foundation' (Galvan, 1992) – even though not a full justification of our trust in the base theory – rooted in our grasp of the notion of truth for T. There are several ways to add a theory of truth to a ground theory; a comprehensive treatment is Halbach (2014).

In this paper we investigate a possible extension of this method. One might see this work as an attempt to climb up Galvan's hierarchy of explanation by resorting to our grasp of 'logical' concepts such as truth itself but also of other modal notions, *in primis* necessity.[2] In other words we investigate the possibility of extending our base theory with 'natural' axioms governing modalities conceived as predicates and not as operators. This line of research is receiving new attention in the recent literature; Quine, Carnap, Montague have all already considered formal treatments of predicative uses of modal notions,[3] but the success of possible world semantics for operator modal logic and the presence of paradoxes in the predicate setting (see §3) have distracted much attention from it.

Halbach et al. (2003) have restored some confidence in the possibility of bridging the gap between modal logics and formal approaches to modal notions conceived as predicates. They have shown that, despite the presence of paradoxes, it is still possible to extend possible-worlds semantics to languages featuring modal predicates at least for some modal frames (cf. §4.1). Halbach and Welch (2009) have even suggested a generalization to arbitrary frames: a variant of their construction will be considered below.

The reader familiar with operator modal logics should not be worried: the predicate approach can be considered a generalization modal logics. Anything that can be said and proved in the operator approach can be mimicked in the predicate setting when suitable restrictions to the predicate language have been performed (see Gupta (1982) and Schweizer (2002)): it is in fact always possible to define an operator via a predicate. What we will say below will be no threat to the usual operator approach; paradoxes arise only when the expressive power of predicates and diagonalization comes into the picture. For the interested reader, Stern (2015) is a thorough and up to date treatment of the current research on

2 We consider truth as a modal notion in the same vein of some medieval logicians such as William of Ockham. See for instance part II of the *Summa Logicae*.

3 See for instance Carnap (1934), Quine (1960), Montague (1970).

syntactical treatments of modalities, including many original contributions by Stern himself.

We end this introductory section with three caveats. First of all we refer to truth, necessity, possibility, etc. as 'logical' notions in a rather liberal sense. Of course we do not advocate the view that the theories considered below amount to 'logics' in the very same sense in which first-order logic is 'logic'; rather we highlight the different possibilities that one faces when extending a given base theory. In this sense we oppose 'logical' principles, such as the ones characterizing *concepts* such as truth and necessity, to ontologically committing 'mathematical' principles, such as set existence assumptions. Furthermore, it is not our intention to suggest a *revision* of modal logics: the predicate approach, in our view, is a framework that naturally that naturally captures the ubiquitous predicative uses of modalities, and it is in this respect an interesting alternative to modal logics or its extensions. Finally, in this work we will only able to partially accomplish the promised ascent given by the combination of alethic modalities. Since the predicate approach to modalities is a lively but young field of research, there is some work required before tackling a fully-fledged proof-theoretic investigation of modal theories: in particular, as we shall see later on, consistency is a highly nontrivial matter.

Plan of The (rest of the) Paper. In §2 we introduce some of the preliminaries needed in the core sections of the paper. Further terminology and notation will be introduced in §4.1. In §3 we focus on some well-known paradoxes of the predicate approach such as Montague's, and on some less well-known antinomies essentially due to the interaction of more than one modal predicate. §4 will be devoted to possible worlds semantics for languages expanding our base language $\mathcal{L}$ with a primitive necessity predicate: we first describe some strategies available to retain a classical interpretation of necessity by restricting the modal space, and then remove these restrictions via a quasi-classical interpretation of necessity based on supervaluations. §5 will finally be devoted to deductive systems: we extend Cantini's theory of truth VF with axioms for necessity and prove its soundness with respect to a multimodal semantics obtained by adapting the semantics given in §4. We conclude in §6 with some comments to the content of the previous sections and sketch some possible extensions.

2 Some preliminaries

Robinson's arithmetic Q is often considered to be the theoretical lower-bound for the derivability of non-intensional independence results such as Gödel's first

incompleteness theorem, Tarski's and Montague's theorems. Let $\mathcal{L} = \{0, \mathsf{S}, +, \times\}$. The axioms of Q are the universal closures of the following formulas:

Q1	$\mathsf{S}x \neq 0$
Q2	$\mathsf{S}x = \mathsf{S}y \rightarrow x = y$
Q3	$x \neq 0 \rightarrow \exists y(x = Sy)$
Q4	$x + 0 = x$
Q5	$x + \mathsf{S}y = \mathsf{S}(x + y)$
Q6	$x \times 0 = 0$
Q7	$x \times \mathsf{S}y = (x \times y) + x$

The axiom Q3 is a weak form of induction and it indispensable to characterize the successor function, as in its absence there may be nonzero natural numbers without a predecessor. Q3 becomes derivable, however, when induction is added to Q.[4]

Peano arithmetic (PA) will play an important role in what follows: it is the result of adding to Q the schema of mathematical induction

(Ind) $$\varphi(0) \wedge \forall x(\varphi(x) \rightarrow \varphi(\mathsf{S}x)) \rightarrow \forall x\, \varphi(x)$$

for all $\mathcal{L}$-formulas $\varphi(v)$ with at most v free.

PA will be the theory formalizing the structure and properties of the bearers of modal ascriptions. We assume a standard arithmetization of the usual primitive recursive syntactic notions and operations of $\mathcal{L}$ and its extensions as it can be found, for instance, in Galvan (1992). In practice, we will work in a definitional extension of PA in which function symbols (e.g. for syntactic operations) for some primitive recursive functions are available. They can however be eliminated in the usual way (see again Galvan (1992)).

As to notational conventions, we only give few instructive examples: $\dot{\neg}x$ stands for the $\mathcal{L}$-term representing in PA the operation of prefixing a negation symbol to x, and similarly $\dot{\mathsf{N}}t$ is the $\mathcal{L}$-term representing the result of prefixing the predicate N of the language $\mathcal{L}_{\mathsf{N}} := \mathcal{L} \cup \{\mathsf{N}\}$ to the object coded by t; $\mathsf{Sent}_{\mathcal{L}}(x)$ is a PA-definable formula representing the primitive recursive set of sentences of $\mathcal{L}$; the $\mathcal{L}$-formula $\mathsf{Bew}_T(x)$ represents the recursively enumerable set of theorems of the recursive theory T; x° stands for the PA-definable evaluation function assigning to each closed term its value. When this is clear from the context, we follow the customary practice and do not distinguish between sentences and their codes.

4 Q is extremely weak. Saul Kripke observed in fact that cardinal numbers are a model of Q, and thus there are entities, such as infinite cardinals, for which $\mathsf{S}x = x$.

We conclude this section by introducing a technical device that is often useful to interpret self-applicable predicates. A sound translation function $\tau\colon \mathcal{L}_{\mathsf{N}} \to \mathcal{L}_1$ for sentences of the form $\mathsf{N}\dot{\mathsf{N}}t$, replacing $\mathsf{N}(\cdot)$ with some $\mathcal{L}_1$-formula $\xi(\cdot)$ should of course yield $\xi(\dot{\tau}\dot{\mathsf{N}}t)$ and not $\xi(\mathsf{N}t)$, where the notation $\dot{\tau}(\cdot)$ is like in the previous paragraph. To achieve the required translation, one may resort to the recursion theorem (Rogers, 1987, §11.2), that yields for any recursive $f(x, y)$ an index e such that $f(e, y) = [e](y)$, where $[\cdot](\cdot)$ is the universal program. If we recursively define a function τ_0 such that, in the relevant case, $\tau_0(x, \mathsf{N}\dot{\mathsf{N}}t) = \xi([x](\dot{\mathsf{N}}t))$, we would then be able to apply the recursion theorem and find an index e for τ_0 such that $[\mathsf{e}](\mathsf{N}\dot{\mathsf{N}}t) = \xi([\mathsf{e}](\dot{\mathsf{N}}t))$. We are done by letting $\tau(x) \equiv [\mathsf{e}](x)$.

3 Montague's paradox and extensions

Paradox is one of the main challenges that the proponent of the predicate approach to modalities has to face. In this section we introduce some paradoxical patterns of reasoning daunting the predicate approach by distinguishing the *unimodal* framework, in which our ground language is extended with only one modality, and a *multimodal* setting, in which more modalities are taken to interact. As it happens, paradox arises in both frameworks.

Montague's paradox is arguably the most fudamental form of paradoxicality in the unimodal setting. The theorem can be stated also in a more general form (Montague, 1974), but here we shall be content with the following.

Lemma 4 (Montague). *Let $T \supseteq \mathsf{Q}$ and assume there is a unary (possibly defined) predicate χ such that, for all $\varphi \in \mathcal{L}_T$:*

(T) $\quad T \vdash \chi^{\ulcorner}\varphi^{\urcorner} \to \varphi$

(NEC) $\quad$ *if* $T \vdash \varphi$, *then* $T \vdash \chi^{\ulcorner}\varphi^{\urcorner}$

Then T is inconsistent.

Proof. By the diagonal lemma, there is a sentence γ of $\mathcal{L}_T$ such that

$$T \vdash \gamma \leftrightarrow \neg\chi^{\ulcorner}\gamma^{\urcorner}$$

Now we reason in T as follows:

$\chi^{\ulcorner}\gamma^{\urcorner} \to \gamma$	(T)
$\chi^{\ulcorner}\gamma^{\urcorner} \to \neg\gamma$	def. of γ

$\neg\chi^{\ulcorner}\gamma^{\urcorner}$	
γ	by def. of γ
$\chi^{\ulcorner}\gamma^{\urcorner}$	(NEC)

qed

It is easy to see why Lemma 4 or variants thereof have led many authors, including Montague, to conclude that virtually no modal reasoning can be carried out in the predicate approach to modality. (T) and (NEC) are in fact basic for our understanding of some modalities, above all *de dicto* necessity.

This is, as we shall see shortly, a rather hasty conclusion. There are many examples of predicative uses of modalities in our philosophical reasoning, including core claims such as 'There are a posteriori necessary truths', or 'Any analytic judgment is necessary', that are most naturally formalized using modal predicates. Some portions of our reasoning with predicative modal asciptions can be rescued from paradox.[5]

One might argue at this stage that, as in the case of the Liar paradox, there is a straightforward way out of paradox given by Tarski's hierarchy of languages. If this is obviously true for the unimodal setting, when we move to languages featuring at least two modalities typing is not a sufficient solution anymore. Halbach (2006), for instance, produced the following, illuminating example involving two modalities M_1 and M_2 that closely resemble truth and necessity.

To formulate Halbach's result, let $T \supseteq \mathsf{Q}$ and expand $\mathcal{L}_T$ with predicates M_1 and M_2; call the resulting language $\mathcal{L}^+$. We say that $\varphi \in \mathcal{L}^+$ does not contain M_i if it does not contain any *used* occurrences of it, but it may contain *mentioned* occurrences.

Proposition 1 (Halbach). *Let T^+ extend T with the axiom schemata*

(1)	$M_1^{\ulcorner}\varphi^{\urcorner} \leftrightarrow \varphi$	*for all $\varphi \in \mathcal{L}^+$ not containing M_1.*
(2)	$M_2^{\ulcorner}\varphi^{\urcorner} \rightarrow \varphi$	*with $\varphi \in \mathcal{L}^+$ not containing M_2*
(3)	$\dfrac{\varphi}{M_2^{\ulcorner}\varphi^{\urcorner}}$	*with $\varphi \in \mathcal{L}^+$ not containing M_2*

Then T^+ is inconsistent.

5 Surely that are ways to strengthen the operator approach and mimic the expressive power of modal predicates, but one can hardly deny that the resulting formalizations will be less natural.

Proof. We reason in T^+:

$\nu \leftrightarrow \neg M_1 \ulcorner M_2 \ulcorner \nu \urcorner \urcorner$	diagonal lemma
$M_2 \ulcorner \nu \urcorner \rightarrow \neg \nu$	by (1)
$M_2 \ulcorner \nu \urcorner \rightarrow \nu$	(2)
$\neg M_2 \ulcorner \nu \urcorner$	
$\neg M_1 \ulcorner M_2 \ulcorner \nu \urcorner \urcorner$	by (1)
ν	def. of ν
$M_2 \ulcorner \nu \urcorner$	(3)

qed

Our emphasis on the use\mention distinction should be now more motivated: the paradox would in fact disappear if instead of sentences of $\mathcal{L}^+$ we had chosen sentences of the ground language $\mathcal{L}_T$, where also mentioned occurrences of the modalities are not allowed.

Proposition 1 is only one of the paradoxes arising from the interaction of modal predicates. For instance, a paradox involving knowledge structurally similar to Proposition 1 can be found in Halbach (2008); Horsten and Leitgeb (2001) also show that seemingly innocuous assumptions on the structure of time lead to the inconsistency of the future. Proposition 1 was preferred to other choices for a simple reason: it suggests that multimodal paradoxes are somewhat harder to eradicate than their unimodal cousins.

Some authors have already set the basis for a systematic study of the multimodal paradoxes and their properties. A promising line of research consists for instance in applying insights from diagonal modal logics to analyse the structure of multimodal paradoxes. The fundamental idea behind this approach is to mimic the expressive power of arithmetic by considering propositional languages expanded with constants for modal ascriptions and a diagonal axiom for each of them. This boost in the expressive power provides enough information to analyse the 'logical' structure multimodal paradoxes. The interested reader may consult Egré (2005) and Fischer and Stern (2015) for further details.

4 Models for necessary truths

What has been said in the last section strongly suggests extra care in handling expansions of $\mathcal{L}$ with modal predicates. Therefore in this section we start more humbly with providing a possible-worlds semantics for the expansion of $\mathcal{L}$ with

a single necessity predicate N. In the next section we will see how to combine a truth predicate with the necessity predicate.

We exclusively focus on *de dicto* necessity, that is we consider only necessity ascriptions that apply to propositions, and not *de re* necessity ascriptions, which attribute a property to an (or possibly more) object by necessity. If the formalization of de dicto necessity as a unary predicate applying to names of sentences seems uncontroversial,[6] there are several options to deal with de re necessity or more generally de re modality. A promising option is to employ a binary predicate applying to unary formulas (playing the role of properties) and sequences of domain objects (variable assignments), mimicking a binary predicate for satisfaction. A careful treatment to de re modality, also in comparison to indexed modalities in modal logic, is deferred to a forthcoming work.

As we have mentioned in the introductory section, there are essentially two ways of constructing a possible world semantics for $\mathcal{L}_{\mathsf{N}} = \mathcal{L} \cup \{\mathsf{N}\}$. One can either consider a specific set of frames and allow for a classical interpretation of N, or instead impose no restrictions to the admissible frames and interpret the necessity predicate in a nonclassical way. We are mostly interested in the latter option, but for the sake of completeness we will also briefly sketch the fundamentals of the former without proofs: some terminology and the core insights of the classical approach will in fact also be useful later on.

4.1 Classical Interpretations of Necessity

We begin with some notions that may sound familiar from operator modal logic, but that it is worth repeating due to the new environment. They will also be useful in later sections.

Definition 1. *Models of $\mathcal{L}_{\mathsf{N}}$ will be pairs $(\mathbb{N}, X)$ where X is the extension of* N. *These pairs are 'worlds' in a possible worlds model. Therefore since we are dealing with standard models of $\mathcal{L}$ only, we may write (w, X) and $(\mathbb{N}, X)$ interchangeably.*

(i) A frame *is a pair (W, R) with $W \neq \varnothing$ and $R \subseteq W \times W$;*

(ii) A possible worlds model is a triple (W, R, V), with (W, R) a frame and V a function from worlds to subsets of $\mathcal{L}_{\mathsf{N}}$ such that for every $w \in W$:

$$V(w) = \{\varphi \in \mathcal{L}_{\mathsf{N}} \mid \forall u(wRu \Rightarrow V(u) \vDash \varphi)\}$$

6 Obviously the controversy may arise at the level of the bearers of modal ascriptions. As usual, the sentence or the proposition are equally good candidates. Following the recent literature we take sentence types to be the bearers of modal ascritpions.

(iii) *A frame (W, R) admits a valuation if and only if there is a V such that (W, R, V) is a possible worlds model.*

We notice that, unsurprisingly, many basic consequences of the definitions carry over in the predicate approach. In particular, we have the standard properties of models of the operator modal logic K.

Lemma 5. *For (W, R, V) a possible worlds model and $w \in W$:*

(i) $(\mathbb{N}, V(w)) \vDash \mathsf{N}\ulcorner\varphi\urcorner$ *if and only if* $(\forall v \in W)(wRv \Rightarrow (\mathbb{N}, V(v)) \vDash \varphi)$

(ii) *if* $(\mathbb{N}, V(v)) \vDash \varphi$ *for all* $v \in W$, $(\mathbb{N}, V(w)) \vDash \mathsf{N}\ulcorner\varphi\urcorner$.

(iii) $(\mathbb{N}, V(w)) \vDash \mathsf{N}\ulcorner\varphi \rightarrow \psi\urcorner \wedge \mathsf{N}\ulcorner\varphi\urcorner \rightarrow \mathsf{N}\ulcorner\psi\urcorner$

Next we finally turn to the differences between the predicate and operator approach. If, given a frame (W, R) and worlds modelling $\mathcal{L}$, we can always construct a model for the language $\mathcal{L} \cup \{\Box\}$ by recursively defining truth for $\mathcal{L}_{\Box}$, the same strategy fails for the language $\mathcal{L}_{\mathsf{N}}$. Only certain frames admit a valuation, due to the paradoxical phenomena considered in the previous section. For instance, Lemma 4 shows that no reflexive frame admits a valuation for the necessity predicate.

One may wonder at this stage whether there are any general criteria to isolate the frames support a valuation. To this end, we introduce new terminology.

Definition 2.

(i) *A (binary) relation R is* converse well-founded *on a set X iff all nonempty $Y \subseteq X$ have an R-maximal element.*

(ii) *A frame (W, R) is converse well-founded iff R is converse well-founded on W.*

(iii) *If (W, R) is a frame and R converse well-founded on W, the* rank *of $w \in W$ is:*

$$\rho(w) := \begin{cases} 0, & \text{if there is no } v \text{ with } wRv \\ \alpha + 1, & \text{if } \exists v(wRv \wedge \rho(v) = \alpha \wedge \forall u(wRu \Rightarrow \rho(u) \leq \alpha)) \end{cases}$$

(iv) *The* converse well-founded *part of $\{v \mid wRv\}$ w.r.t. W is the largest R-upwards closed $X \subseteq W$ such that R^{-1} is well-founded on X.*

(v) *The* rank *of a converse ill-founded world is the rank of its converse well-founded part.*

If $\rho(w) = 0$, we say that w is a *dead end*.

Depending on the frames considered, it is possible to impose sufficient conditions on the existence of valuations. Converse well-foundedness is one of them. By transfinite induction on the rank of $w \in W$ in a converse well-founded

frame (W, R) one defines the valuation

$$V(w) := \{\varphi \in \mathsf{Sent}_{\mathcal{L}_\mathsf{N}} \mid \forall v(wRv \Rightarrow (\mathbb{N}, V(v)) \vDash \varphi)\} \tag{4}$$

The crucial point is that if R is converse well-founded, we have the following picture for any $w \in W$,

In other words, from any w it is alway possible to reach a dead end, u_1 in this case, in finitely many steps. The valuation defined by (4) is thus unique, yielding

Proposition 2 (Gupta & Belnap). *If (W, R) is converse well-founded, it admits a unique valuation.*

For frames containing converse ill-founded worlds, it is also possible to find a valuation under certain circumstances. To see this, let us consider the operator $\Phi\colon \mathcal{P}(\omega) \to \mathcal{P}(\omega)$

$$\Phi(X) := X \cap \{\varphi \in \mathsf{Sent}_{\mathcal{L}_\mathsf{N}} \mid (\mathbb{N}, X) \vDash \varphi\}$$

$\Phi(\cdot)$ is a decreasing and anti-monotone operator (i.e. $\Phi(Y) \subseteq Y$ for all Y and $\alpha \leq \beta$ entails that $\Phi^\alpha(Y) \supseteq \Phi^\beta(Y)$). Therefore, if one starts with $\Phi^0(\mathcal{L}_\mathsf{N}) := \Phi(\mathcal{L}_\mathsf{N})$ and iterates the application along an ordinal path – taking intersections at limit stages – one reaches a fixed point with associated a closure ordinal, that is a stage κ in which $\Phi^\kappa(\mathcal{L}_\mathsf{N}) = \Phi^\beta(\mathcal{L}_\mathsf{N})$ for all $\beta \geq \kappa$. The closure ordinal of $\Phi(\cdot)$ has also been computed by Halbach et al. (2003) as the least α such that the corresponding level of the constructible hierarchy L_α possesses a Σ_1-elementary end extension (Halbach et al., 2003, Prop. 21). In particular, we have $\kappa > \omega_1^{\mathsf{CK}}$, the first nonrecursive ordinal.

If a frame (W, R) is transitive and has converse ill-founded worlds w, the fixed point $\Phi^\kappa(\mathcal{L}_\mathsf{N})$ can always be used as valuation when the rank of w is greater than or equal to κ. That is

Proposition 3 (Halbach et al. (2003)). *If (W, R) is transitive and the rank of its converse ill-founded worlds is not smaller than κ, then (W, R) supports a valuation.*

The closure ordinal κ is also useful to impose necessary conditions on the existence of valuations in transitive frames. Let A be the class of admissible ordinals (without ω) with limits (see for instance Devlin (1984)).

Proposition 4 (Halbach et al. (2003)). *If (W, R) is transitive and admits a valuation, then for a converse ill-founded world $w \in W$, either $\rho(w) \in \mathsf{A}$ or $\rho(w) \geq \kappa$.*

Proposition 4 tells us that if (W, R, V) is a possible worlds model and R is converse ill-founded, then there will always be, for $w \in W$, an initial well-ordered portion of 'rank' $\alpha \in \mathsf{A}$ or greater-equal than κ. This means that frames (W, R) whose worlds have rank less than the first admissible ordinal ω_1^{CK} admit a valuation if and only if R is converse well-founded. Moreover, Proposition 4 can be generalized to non transitive frames, if we focus on the transitive closure of the accessibility relation.

In this brief overview our main intention was to highlight a fundamental fact: if one is interested in a classical interpretation of the necessity predicate, there are strong limitations one has to face. Again this is not a problem for the predicate approach if opposed to the operator approach, as we have already mentioned that the operator language can be straightforwardly translated in the predicate language. The problem is internal to the predicate approach. There is in fact an alternative to the classical approach sketched in this section: we can preserve the generality of the possible worlds semantics for operator modal logics if we move to a nonclassical setting.

4.2 Arbitrary Frames: Supervaluations

In this section we present a method for constructing possible worlds models for arbitrary frames (W, R). As before, worlds $w \in W$ are standard models of the ground language $\mathcal{L}$. The strategy is reminiscent of Kripke's fixed-point construction (Kripke, 1975), which can be also seen as a method for generating models for $\mathcal{L}_{\mathsf{N}}$ in a reflexive frame $(\{w\}, R)$. To produce models for arbitrary W, one has to generalize Kripke's construction.

Halbach and Welch (2009) have proposed a similar generalization of Kripke's theory based on the Strong Kleene evaluation schema. We explore an alternative option and employ the supervaluational scheme introduced by Van Fraassen (1966). We will highlight some nice features of supervaluations as opposed to the Strong Kleene approach after introducing few definitions and some of their consequences.

As before, $(w, F(X))$ will denote a model of $\mathcal{L}_{\mathsf{N}}$ in which w specifies the standard model of the ground language but X is now an evaluation function $F\colon W \to (\mathsf{Sent}_{\mathcal{L}_{\mathsf{N}}} \times \mathsf{Sent}_{\mathcal{L}_{\mathsf{N}}})$: at each world $w \in W$ it assigns *disjoint* extensions and antiextension to N. We also define an ordering $\preceq$ between evaluation functions such that $F \preceq G$ if, at any $w \in W$, $F(w)^+ \subseteq G(w)^+$ and $F(w)^- \subseteq G(w)^-$. We define a

(binary) relation $\models_{\mathrm{vf0}}$ linking pairs (w, X) and $\mathcal{L}_{\mathsf{N}}$-sentences φ:[7]

$$(w, F(w)) \models_{\mathrm{vf0}} \varphi :\Leftrightarrow (\forall G)(F \preceq G \wedge G(w)^{+} \subseteq \omega \setminus F(w)^{-} \Rightarrow (w, G(w)^{+}) \models \varphi)$$

The condition $G(w)^{+} \subseteq \omega \setminus F(w)^{-}$, together with the disjointness of extension and antiextension, will force consistent fixed points as extensions of the necessity predicate; that is, at any world w for no $\varphi \in \mathcal{L}_{\mathsf{N}}$, $\neg\varphi$ and φ will be in the ultimate extension of the necessity predicate. The relation $\models_{\mathrm{vf0}}$ extends the standard supervaluational picture according to which truth is satisfaction in all candidate extensions of a starting set; here we have merely generalized this picture to many worlds.[8]

To assign a suitable interpretation to the necessity predicate, we consider a variant of the strategy adopted by Halbach and Welch (2009) and impose further conditions on evaluation functions. We let EV be the set of such evaluations:

Definition 3. *The operator $\Delta\colon \mathsf{EV} \to \mathsf{EV}$, at each $w \in W$, is such that:*

$$(\Delta(F))(w)^{+} := \{\varphi \mid \forall v(wRv \Rightarrow (v, F(v)) \models_{\mathrm{vf0}} \varphi)\}$$
$$(\Delta(F))(w)^{-} := \{\varphi \mid \exists v(wRv \wedge (v, F(v)) \models_{\mathrm{vf0}} \neg\varphi)\}$$

The following is an immediate corollary of the definitions.

Corollary 3. *The operator Δ is monotone with respect to $\preceq$, that is, for all $w \in W$,*

$$F \preceq G \Rightarrow (\Delta(F))(w) \preceq (\Delta(G))(w)$$

The monotonicity of Δ implies the existence of fixed points. This follows from abstract cardinality considerations (Moschovakis, 1974, Thm. 1.A.1). As before, we may track the applications of Δ on an ordinal path using suitable indices. In other words $\Delta^{\alpha}(F)(w)$ denotes the α^{th} application of Δ to the starting evaluation function F at a world w, taking unions at limit stages. A fixed point of Δ will thus be an ordinal κ such that $\Delta^{\kappa}(F)(w) = \Delta^{\beta}(F)(w)$ for all $\beta \geq \kappa$.

By reflecting on the properties of $\Delta(\cdot)$, we have

Proposition 5. *If F is a fixed point of Δ, for all $\varphi \in \mathcal{L}_{\mathsf{N}}$ and frames (W, R) with $w \in W$:*

(5) $(w, F(w)) \models_{\mathrm{vf0}} \mathsf{N}\ulcorner\varphi\urcorner \Leftrightarrow$ *for all v, if wRv, then $(v, F(v)) \models_{\mathrm{vf0}} \varphi$*

(6) $(w, F(w)) \models_{\mathrm{vf0}} \neg\mathsf{N}\ulcorner\varphi\urcorner \Leftrightarrow$ *exists a v with wRv and $(v, F(v)) \models_{\mathrm{vf0}} \neg\varphi$*

7 This is in a sense a simplifying choice: we dispense with variable assignments as we assume that we have constant domains and fixed names for all objects at every $w \in W$.

8 There are other possible choices of the evaluational scheme, still in the supervaluational spirit. See Burgess (1987) or Fischer et al. (2015).

Proof.

Ad (5). ($\Rightarrow$) If $(w, F(w)) \models_{\mathrm{vf0}} \mathsf{N}\ulcorner\varphi\urcorner$, then for all evaluations $G \succeq F$, including F itself: if $G(w)^+ \subseteq \omega \setminus F(w)^-$, then $(w, G(w)^+) \models \mathsf{N}\ulcorner\varphi\urcorner$. Therefore, $\varphi \in F(w)^+$. Since F is a fixed point of Δ, $F(w)^+ = (\Delta(F))(w)^+$, and $\varphi \in (\Delta(F))(w)^+$, that is

$$\forall v(wRv \Rightarrow (v, F(v)) \models_{\mathrm{vf0}} \varphi)$$

($\Leftarrow$) If for all v with wRv, $(v, F(v)) \models_{\mathrm{vf0}} \varphi$, by definition of Δ also $\varphi \in (\Delta(F))(w)^+$. Again by the fixed point property, $\varphi \in F(w)^+$. This means that for all evaluations $G \succeq F$, and a fortiori the ones in which $G(w)^+ \subseteq \omega \setminus F(w)^-$, $\varphi \in G(w)^+$. By the classical satisfaction relation, $(w, G(w)^+) \models \mathsf{N}\ulcorner\varphi\urcorner$. By definition of $\models_{\mathrm{vf0}}$ we finally obtain

$$(w, F(w)) \models_{\mathrm{vf0}} \mathsf{N}\ulcorner\varphi\urcorner$$

Ad (6). ($\Rightarrow$): If $(w, F(w)) \models_{\mathrm{vf0}} \neg\mathsf{N}\ulcorner\varphi\urcorner$ the for all $G \succeq F$ and $G(w)^+ \subseteq \omega \setminus F(w)^-$, $\varphi \notin G(w)^+$. A fortiori, $\varphi \notin \omega \setminus \mathsf{NSent}_{\mathcal{L}_\mathsf{N}} \setminus F(w)^-$ where $\mathsf{NSent}_{\mathcal{L}_\mathsf{N}}$ is the set of numbers that are not $\mathcal{L}_\mathsf{N}$-sentences; therefore $\varphi \in F(w)^- = (\Delta(F))(w)^-$.

($\Leftarrow$). If $\exists v(wRv \wedge (v, F(v)) \models_{\mathrm{vf0}} \neg\varphi)$, then $\varphi \in (\Delta(F))(w)^- = F(w)^-$. Therefore $(w, G(w)^+) \models \neg\mathsf{N}\ulcorner\varphi\urcorner$ for any $G \succeq F$ and $G(w)^+ \subseteq \omega \setminus F(w)^-$, that is

$$(w, F(w)) \models_{\mathrm{vf0}} \neg\mathsf{N}\ulcorner\varphi\urcorner$$

qed

The *minimal* fixed point $\mathfrak{I}_\Delta$ is obtained by closing the empty evaluation under Δ at any world $w \in W$, and it is the minimal fixed point that we now examine to highlight some nice features of the supervaluationist approach to necessity.

Let us call the *contingency teller* the sentence μ such that

$$\mathsf{Q} \vdash \mu \leftrightarrow \neg\mathsf{N}\ulcorner\mu\urcorner$$

Montague's paradox rules out reflexive frames in the classical setting. The contingency teller played an important role in the proof of Lemma 4. To see how the nonclassical setting helps in dealing with paradoxes, we now show that in the new setting μ will be 'gappy', that is neither necessary or contingent, in the minimal fixed point.

Lemma 6. *Let (W, R) be a frame. The contingency teller is neither in $\mathfrak{I}^+_\Delta(w)$ nor in $\mathfrak{I}^-_\Delta(w)$ for any $w \in W$.*

Proof. We prove the claim by induction on the construction of the minimal fixed point of Δ.

At stage $\Delta^0(\varnothing)(w) := (\varnothing, \varnothing)$, the claim is trivially satisfied.

At successor stages $\alpha+1$, if $\mu \in \mathcal{I}_\Delta^{\alpha+1}(w)^+$, then

$$\forall v(wRv \Rightarrow (v, \mathcal{I}_\Delta^\alpha(v)) \vDash_{\mathrm{vf0}} \neg \mathsf{N}\ulcorner\mu\urcorner) \tag{7}$$

That is, $\mu \notin G(v)^+ \subseteq \omega \setminus \mathcal{I}_\Delta^\alpha(v)^-$ for all suitable G, including $\omega \setminus \mathsf{NSent}_{\mathcal{L}_\mathsf{N}} \setminus \mathcal{I}_\Delta^\alpha(v)^-$. Therefore $\varphi \in \mathcal{I}_\Delta^\alpha(v)^-$, quod non by induction hypothesis.

If, by contrast, $\mu \in I_\Delta^{\alpha+1}(w)^-$, there will be, for all extensions G of $\mathcal{I}_\Delta^\alpha$, $(v, G(v)^+) \vDash \mathsf{N}\ulcorner\mu\urcorner$ at some accessible v. Thus $\mu \in I_\Delta^\alpha(v)^+$, again contradicting the induction hypothesis.

Finally, if $\mu \in \mathcal{I}_\Delta^\lambda$ for a limit λ, the claim follows from the previous steps by definition of $\Delta(\cdot)$. qed

By suitably adapting the argument of Lemma 6, one easily shows that $\neg\mu$ cannot be in $\mathcal{I}_\Delta$. Moreover, a generalization of this arguments shows that there are *consistent* fixed points of Δ.

As we have already observed, the operator $\Delta(\cdot)$ compares to the operator based on the Strong Kleene evaluation schema considered in Halbach and Welch (2009). It is well-known since Kripke (1975) that the Strong Kleene schema yields an attractive picture of self-applicable truth predicate. Above all, it yields a compositional semantics, e.g. $A \vee B$ is $\mathrm{true}_{\mathrm{sk}}$ if and only if A is $\mathrm{true}_{\mathrm{sk}}$ or B is $\mathrm{true}_{\mathrm{sk}}$ with A, B sentences of a base language such as $\mathcal{L}$ plus a primitive truth predicate.

If necessity and not truth simpliciter is at stake, one may argue that compositionality is not as important as, for instance, establishing the necessity of all laws of classical logic; so $A \vee \neg A$ should be necessary even though we do not have the resources to find out whether A or its negation are true. The following results show that the supervaluationist approach captures, in the predicate approach, the picture of necessity just sketched.

Proposition 6.

(i) All logical laws, including the laws of the conditional (e.g. $\varphi \to \varphi$ for $\varphi \in \mathcal{L}_\mathsf{N}$) valid in $\mathcal{I}_\Delta$ (i.e. in $\mathcal{I}_\Delta^+(w)$ at any w);

(ii) Let PAN *simply* PA *formulated in $\mathcal{L}_\mathsf{N}$. All theorems of* PAN *are valid in $\mathcal{I}_\Delta$.*

Proof. In both cases one reflects on the definition of $\Delta(\cdot)$. At stage 1 of the construction of $\mathcal{I}_\Delta(w)$ for arbitrary w, we have

$$(\Delta^1(\varnothing, \varnothing))(w) = \left\langle \begin{array}{l} \{\varphi \mid \forall v(wRv \Rightarrow (v, (\varnothing, \varnothing)) \vDash_{\mathrm{vf0}} \varphi)\}, \\ \{\varphi \mid \exists v(wRv \& (v, (\varnothing, \varnothing)) \vDash_{\mathrm{vf0}} \neg\varphi)\} \end{array} \right\rangle$$

By definition of $\vDash_{\mathrm{vf0}}$, therefore, all theorems of first-order logic and of PAN will get in $(\Delta^1(\varnothing, \varnothing))(w)^+$; therefore by the monotonicity of Δ also in $\mathcal{I}_\Delta(w)^+$. qed

As a corollary, $\mathsf{N}\ulcorner\mu \vee \neg\mu\urcorner$ will be valid in the fixed point $\mathcal{J}_\Delta$ at any world, although μ, as we have seen already, will not be in any fixed point. In addition, also bi-conditionals containing gappy sentences such as $\mu \leftrightarrow \neg\mathsf{N}\ulcorner\mu\urcorner$ will be in the fixed point.

We have thus seen that there are ways to overcome the paradoxes of the predicate approach and capture predicative uses of necessity by providing *models* for the base language expanded with a predicate for necessity. In the next section we consider some strategies to formulate deductive systems inspired to the semantic construction just given.

5 A system for truth and necessity

In this section we move the first steps into combining truth and necessity. We introduce a modal version of Cantini's VF Cantini (1990) and prove its soundness with respect to a modification of the semantics given in the previous section.

5.1 The theory VF

VF is the theory capturing the properties of a self-applicable (type-free) truth predicate interpreted according to a suitable modification of the operator Δ introduced above:

$$(\mathbb{N}, X) \vDash_{\mathsf{vf}} \varphi :\Leftrightarrow \forall S(X \subseteq S \wedge \mathsf{con}^*(S) \Rightarrow (\mathbb{N}, S) \vDash \varphi)$$

Here we have dropped the antiextension and we deal only with consistent candidate extension: in particular $\mathsf{con}^*(G(w))$ expresses that $G(w)$ does not contain negations of sentences in $F(w)$; to avoid triviality, *only consistent* starting evaluations $F(w)$ are allowed. Let $\mathcal{L}_\mathsf{T}$ be the language $\mathcal{L}$ expanded with a unary truth predicate T. We call the new operator $\Theta\colon \mathcal{P}(\omega) \to \mathcal{P}(\omega)$:

$$\Theta(X) := \{\varphi \mid (\mathbb{N}, X) \vDash_{\mathsf{vf}} \varphi\}$$

By only a slight modifications of the arguments already given there, we notice that Θ is monotone and thus it has fixed points. We define by transfinite induction

$$\begin{aligned} \mathcal{J}_\Theta^0 &:= \varnothing \\ \mathcal{J}_\Theta^{\alpha+1} &:= \Theta(\mathcal{J}_\Theta^\alpha) \\ \mathcal{J}_\Theta^\lambda &:= \bigcup_{\beta<\lambda} \mathcal{J}_\Theta^\beta \end{aligned}$$

The minimal fixed point $\mathcal{J}_\Theta$ is simply $\mathcal{J}_\Theta^\kappa$, where κ is the closure ordinal for Θ.

Cantini (1990) introduced a deductive system that is sound with respect to fixed points of Θ. It is called VF from 'Van Frassen', who first introduced the supervaluational scheme to analyse vague predicates.

Definition 4. VF *is formulated in* $\mathcal{L}_{\mathsf{T}}$. *Its axioms are all axioms of* PAT *(i.e.* PA *formulated in* $\mathcal{L}_{\mathsf{T}}$*) and the following:*

(VF1) $\forall \vec{x}(\mathsf{T}^{\ulcorner}\varphi(\dot{\vec{x}})^{\urcorner} \rightarrow \varphi(\vec{x}))$ *for all* $\varphi \in \mathcal{L}_{\mathsf{T}}$

(VF2) $\forall s,t\big((\mathsf{T}(s \doteq t) \leftrightarrow s^{\circ} = t^{\circ}) \wedge (\mathsf{T}(s \dot{\neq} t) \leftrightarrow s^{\circ} \neq t^{\circ})\big)$

(VF3) $\forall x(\mathsf{Ax}_{\mathsf{PAT}}(x) \rightarrow \mathsf{T}x)$

(VF4) $\forall v \forall x \forall t(\mathsf{T}x(t/v) \rightarrow \mathsf{T}\dot{\forall} vx)$

(VF5) $\forall t(\mathsf{T}t^{\circ} \rightarrow \mathsf{TT}\dot{t})$

(VF6) $\forall s(\mathsf{Sent}_{\mathcal{L}_{\mathsf{T}}}(s^{\circ}) \wedge \mathsf{T}\dot{\neg}\mathsf{T}x \rightarrow \mathsf{T}\dot{\neg}s^{\circ})$

(VF7) $\forall x,y(\mathsf{Sent}_{\mathcal{L}_{\mathsf{T}}}(x \dot{\rightarrow} y) \rightarrow (\mathsf{T}(x \dot{\rightarrow} y) \rightarrow \mathsf{T}x \rightarrow \mathsf{T}y))$

(VF8) $\forall x(\mathsf{T}^{\ulcorner}\mathsf{T}\dot{x} \rightarrow \neg\mathsf{T}\dot{\neg}\dot{x}^{\urcorner})$

(VF9) $\mathsf{T}^{\ulcorner}\mathsf{T}\dot{x} \rightarrow \mathsf{Sent}_{\mathcal{L}_{\mathsf{T}}}(\dot{x})^{\urcorner}$

It is a routine task to check, by induction on the length of the derivation in VF, that

Proposition 7 (Cantini (1990), Prop. 3.4)**.** *If* X *is a fixed point of* Θ*, then* $(\mathbb{N}, X) \vDash$ VF.

5.2 Modal extensions of VF

To introduce a modal extension of VF, we consider a variant of the strategy adopted by Stern (2014) to extend the Kripke-Feferman system KF.[9]

We first introduce predicative counterparts of the well-known modal principles (*T*), (4) and (*E*) formulated in the language $\mathcal{L}_{\mathsf{TN}} := \mathcal{L} \cup \{\mathsf{T}\} \cup \{\mathsf{N}\}$:

(T) $\forall x(\mathsf{Sent}_{\mathcal{L}_{\mathsf{TN}}} \wedge \mathsf{N}x \rightarrow \mathsf{T}x)$

(4) $\forall t\,(\mathsf{TN}\dot{t} \rightarrow \mathsf{NN}\dot{t})$

(E) $\forall t\,(\mathsf{T}\dot{\neg}\mathsf{N}\dot{t} \rightarrow \mathsf{N}\dot{\neg}\mathsf{N}\dot{t})$

As it is well-known from operator modal logic, (T) forces reflexive frames, (4) transitive frames, and (E) Euclidean frames. (T) in combination with (E) suffice to force frames based on an equivalence relation.

9 See again Halbach (2014) for a thorough introduction to KF.

We finally define the theory MVF. The theory PATN is, as one might expect, simply PA formulated in $\mathcal{L}_{\mathsf{TN}}$.

Definition 5 (Modal VF). MVF *is the theory in* $\mathcal{L}_{\mathsf{TN}}$ *whose axioms are (i) the axioms of* PATN, *(ii)* VF *formulated in* $\mathcal{L}_{\mathsf{TN}}$, *(iii) the following sentences and rules:*

(T-in) $\forall t\,(\mathsf{N}t^\circ \rightarrow \mathsf{TN}\dot{t})$

(BF) $\forall v \forall x(\mathsf{Sent}_{\mathcal{L}_{\mathsf{TN}}}(\dot{\forall} v x) \rightarrow (\forall t\, \mathsf{N}x(t/v) \rightarrow \mathsf{N}\dot{\forall} v x))$

(Rig1) $\forall s, t \forall v \forall x(\mathsf{Sent}_{\mathcal{L}_{\mathsf{TN}}}(\dot{\forall} v x) \rightarrow (s^\circ = t^\circ \rightarrow (\mathsf{N}x(s/v) \leftrightarrow \mathsf{N}x(t/v))))$

(Rig2) $\forall s \forall t(s^\circ \neq t^\circ \rightarrow \mathsf{N}(s \dot{\neq} t))$

(K) $\forall x \forall y(\mathsf{Sent}_{\mathcal{L}_{\mathsf{TN}}}(x \dot{\rightarrow} y) \rightarrow (\mathsf{N}(x \dot{\rightarrow} y) \rightarrow (\mathsf{N}x \rightarrow \mathsf{N}y)))$

$$(\mathsf{Nec})\ \frac{\mathsf{T}\ulcorner\varphi\urcorner}{\mathsf{N}\ulcorner\varphi\urcorner} \qquad \text{for all } \varphi \in \mathcal{L}_{\mathsf{TN}}$$

It is worth emphasising that the axiom VF3 declaring the truth of all axioms of PAT now becomes

$$\forall x(\mathsf{Ax}_{\mathsf{PATN}}(x) \rightarrow \mathsf{T}x) \tag{8}$$

As before, by a straightforward induction, we can conclude that all theorems of PATN are true. This includes, for instance, all instances of excluded middle in the language $\mathcal{L}_{\mathsf{TN}}$.

As we have seen in the case of the paradoxes of interaction, eradicating inconsistencies in the multimodal framework is more difficult than in the unimodal setting. Therefore we first ensure that MVF is consistent by reducing its consistency to the consistency of VF. This will also give an upper bound to the proof-theoretic strength of MVF that will be discussed further in the concluding section. The lower bound is clear as VF is contained in MVF.

Proposition 8. MFV *is consistent, if* VF *is.*

Proof. We define the primitive recursive translation $\tau : \mathcal{L}_{\mathsf{TN}} \rightarrow \mathcal{L}_{\mathsf{T}}$ as follows, using the remarks at the end of §2:

$\tau(\varphi) := \varphi$ for $\varphi \in \mathcal{L}$ that is, φ arithmetical

$\tau(\mathsf{T}\ulcorner\varphi\urcorner) := \mathsf{T}\dot{\tau}\ulcorner\varphi\urcorner$

$\tau(\mathsf{N}\ulcorner\varphi\urcorner) := \mathsf{T}\dot{\tau}\ulcorner\varphi\urcorner$

τ commutes with propositional connectives and quantifiers

In essence, the translation just maps necessity into truth. It is easy to verify that the translations of all axioms of MVF are provable in VF. qed

Cantini (1990) showed that VF proves the same arithmetical sentences as the theory ID_1 of elementary positive inductive definitions (see Pohlers (2009)). Proposition 8, therefore, yields the following analysis of MVF.

Corollary 4. MVF *proves the same arithmetical sentences as* ID_1.

5.3 Semantics and Soundness

Proposition 8 gives us a consistency proof for MVF and indirectly a semantics for it; in any model of VF we can construct an internal model of MVF. This does not mean, however, that there are 'nice' models of MVF: in this section we show that there are 'standard models' of MVF obtained by generalizing in a rather natural way the intended models of VF.

We now adapt the semantics given in §4.2 to the multimodal framework. Given a frame $\mathcal{F}$, a model of the language $\mathcal{L}_{\mathsf{TN}}$ at a world $w \in W$ (again we think of $w \in W$ as standard models of $\mathcal{L}$) will be a triple $\mathcal{M}_w := (w, E(w), N_{E(w)})$, where $E\colon W \to \mathcal{P}(\omega)$ is a function assigning to each w an extension of the truth predicate. From this extension one standardly defines an extension of the necessity predicate $N_{E(w)}$ by taking the intersection of the set of truths at all accessible worlds:

$$N_{E(w)} := \{\varphi \in \mathcal{L}_{\mathsf{TN}} \mid \forall v(wRv \Rightarrow \varphi \in E(v))\}$$

The set of truths at all accessible worlds will then be defined using again the supervaluational scheme, but this time to define the extension of the truth predicate and not of the necessity predicate directly. Notice now that we can drop the superscript $^+$ or $^-$ as we are only assigning candidate extensions and not also an antiextension to the predicate. As before, let $\preceq_1$ an ordering of the evaluation functions defined by: $E_0 \preceq_1 E_1$ if and only if for all $w \in W$, $E_0(w) \subseteq E_1(w)$. Therefore we set, for $\varphi \in \mathcal{L}_{\mathsf{TN}}$:

$$\begin{aligned}(w, F(w), N_{F(w)}) \vDash_{\mathsf{vf1}} \varphi :\Leftrightarrow \\ (\forall G_1 \succeq F)(\mathsf{con}^*(G(w)) \Rightarrow (w, G(w), N_{G(w)}) \vDash \varphi)\end{aligned}$$

with $N_{X(w)}$ as above. With $\mathsf{con}^*(G(w))$ we mean again that $G(w)$ does not contain negations of sentences in $F(w)$; as above, to avoid triviality, we consider *only consistent* starting evaluations $F(w)$. The operator $H^{\mathcal{F}}$ on evaluation functions,

relative to a frame $\mathcal{F}$, is defined as

$$(H^{\mathcal{F}}(E))(w) := \{\varphi \in \mathcal{L}_{\mathsf{TN}} \mid (w, E(w), N_{E(w)}) \vDash_{\mathsf{vf1}} \varphi\}$$

The following is an immediate consequence of the definitions.

Lemma 7. *The operator $H^{\mathcal{F}}$ is monotone with respect to $\preceq_1$.*

As before, monotonicity implies the existence of fixed points, that is evaluations such that $H^{\mathcal{F}}(E) = E$. In a fixed point of $H^{\mathcal{F}}(E)$, therefore, for any $\varphi \in \mathcal{L}_{\mathsf{TN}}$, and any world in $\mathcal{F}$,

(9) $(w, E(w), N_{E(w)}) \vDash_{\mathsf{vf1}} \mathsf{T}\ulcorner\varphi\urcorner \Leftrightarrow (\forall G\ {}_1\!\succeq E)(\mathsf{con}^*(G(w)) \Rightarrow (w, G(w), N_{G(w)}) \vDash \varphi)$

(10) $(w, E(w), N_{E(w)}) \vDash_{\mathsf{vf1}} \mathsf{N}\ulcorner\varphi\urcorner \Leftrightarrow \forall v(wRv \Rightarrow \varphi \in E(v))$

Closing the empty evaluation function under iterated applications of $H^{\mathcal{F}}$ along an ordinal path, we reach the minimal fixed point of $\mathcal{J}_{H^{\mathcal{F}}}$ of $H^{\mathcal{F}}$. MVF, however, is not only sound with respect to the minimal fixed point, but it is satisfied by *all* fixed points of $H^{\mathcal{F}}$.

Proposition 9. *Let $\mathcal{F} = (W, R)$ be a frame and R an equivalence relation. If $H^{\mathcal{F}}(E) = E$, then $(w, E(w), N_{E(w)}) \vDash \mathsf{MVF}$ for any $w \in W$.*

Proof. For the PATN axioms and the axioms of VF one merely adapts Cantini's proof. We consider the genuinely modal axioms of MVF.

Ad (T). If $(w, E(w), N_{E(w)}) \vDash \mathsf{N}\ulcorner\varphi\urcorner$, then $\varphi \in E(v)$ for all v accessible from w. By reflexivity, $\varphi \in E(w)$.

Ad (4). Without loss of generality, we can reason about (the code of) a sentence φ. Let us assume $\mathsf{N}\ulcorner\varphi\urcorner \in E(w)$. This entails:

(11) $$(\forall G\ {}_1\!\succeq E)(\mathsf{con}^*(G(w)) \Rightarrow (w, G(w), N_{G(w)}) \vDash \mathsf{N}\ulcorner\varphi\urcorner)$$

Therefore, for all extended evaluations G, $\varphi \in N_{G(w)}$, that is

(12) $$\forall v(wRv \Rightarrow \varphi \in G(v))$$

Now assume $(w, E(w), N_{E(w)}) \vDash \neg\mathsf{N}\mathsf{N}\ulcorner\varphi\urcorner$. There is then a v with wRv and $\mathsf{N}\ulcorner\varphi\urcorner \notin E(v)$, that is

(13) $$\exists v(wRv \wedge (\exists G\ {}_1\!\succeq E)(\mathsf{con}^*(G(v)) \wedge (v, G(v), N_{G(v)}) \vDash \neg\mathsf{N}\ulcorner\varphi\urcorner)$$

By iterating the same reasoning for $N_{G(v)}$, we find

$$\exists v_0(vRv_0 \wedge \varphi \notin G(v_0)) \tag{14}$$

By transitivity, (14) contradicts (12).

The reasoning for (E) is similar to the previous case. So we consider (T-in). If $t^\circ \in N_{E(w)}$, then $t^\circ \in E(v)$ for all v accessible from w. If $\mathsf{N}t \notin E(w)$, then

$$(\exists G\,{}_1\!\succeq E)(\mathsf{con}^*(G(w)) \wedge (w, G(w), N_{G(w)}) \not\models \mathsf{N}t) \tag{15}$$

Again we arrive at

$$\exists v(wRv \wedge t^\circ \notin G(v)) \tag{16}$$

But (16) contradicts $G\,{}_1\!\succeq E$ and $t^\circ \in E(v)$.

We conclude the proof by considering the case of (K). Let us assume $\varphi \to \psi \in N_{E(w)}$ and $\varphi \in N_{E(w)}$; that is

$$\forall v(wRv \Rightarrow \varphi \to \psi \in E(v) \wedge \varphi \in E(v)) \tag{17}$$

The definition of classical satisfaction yields the desired result. qed

In the following, concluding section we elaborate on the results just presented and on the possibility of further work.

6 Conclusion

In the introduction we sketched a project: the formulation of natural systems of interacting modalities extending a some trustworthy theory of modal ascriptions. The 'naturality' criterion imposed on the project has been spelled out in terms of a possible worlds semantics for modal predicates. We have seen that this is a nontrivial matter; paradoxes threaten our predicative uses of modal notions and impose severe restrictions to the space of models of the corresponding languages.

Despite these difficulties, we formulated a system of truth and necessity MVF that adapts the motivation behind Cantini's VF to the new language and that is sound with respect to a rather natural semantics for truth and necessity. In §1 we have recalled Sergio Galvan's idea of a hierarchy of theories that are capable of making explicit our trust in the theories lower down, starting with our preferred theory of the bearers of modal ascriptions (e.g. PA). By Proposition 8, MVF will prove the same arithmetical sentences as ID_1: to give an idea of how this relates to what we called Galvan's hierarchy, we notice that, for instance, the results

of iterating ACA for all ordinals $< \Gamma_0$, the so-called Feferman-Schütte ordinal, is reducible to ID_1. Reflecting on the fact that ACA was already sufficient to formalize the metatheoretic soundness proof for PA, this gives us an idea of how far MVF takes us.

There are however, also some limitations to the success of the strategy of combining modalities. In the first place it does not seem to be possible to achieve a full adequacy result for MVF, exactly as in the case of VF. More precisely, Fischer et al. (2015) have proposed the following criterion of adequacy for systems of truth: a system T is ω-categorical if

$$(\mathbb{N}, S) \vDash T \Leftrightarrow S \in \mathfrak{M} \tag{18}$$

where $\mathfrak{M}$ is a class of acceptable interpretations of the truth predicate given by some semantic theory of truth. Proposition 7 tells us that the right-to-left direction holds for VF. Fischer et al. (2015), by adapting a previous result of Philip Welch, show in fact that the left-to-right direction of (18) cannot be achieved if $\mathfrak{M}$ is the class of supervaluational fixed-points: in nuce, the property of being a supervaluational fixed point is Π^1_1-complete, if (18) held, we would have a Σ^1_1-definition of a supervaluational fixed point. This shows also that such categoricity result cannot be achieved for MVF.

In addition, we know already that Proposition 8 shows that the proof-theoretic strength of MVF does not exceed the one of VF. This is in some sense good news; at the conceptual level we might even welcome the fact that the notion of necessity axiomatized by MVF is in continuity with the corresponding notion of truth and allows for a 'collapse' of necessity into truth in the one world reading. However, it is also true that the interaction of truth and necessity enriches our expressive capability and we would like our modal theory to exceed the strength of the truth theory on which it is based.

These drawbacks of the strategy proposed in this paper may be nonetheless good guiding principles for adopting different strategies to combine truth and necessity: one might for instance follow McGee (1991) and Halbach (2001) and consider necessity as provability in a suitable system. We defer a treatment of this option to further work, but it will most likely lead to a considerable increase in proof-theoretic strength.

Edmund Runggaldier

The Wittgensteinian and the ontological (3-dimensional) reaction to the naturalistic challenge

Abstract: Starting with a trustworthy theory T, Galvan (1992) suggests to read off, from the usual hierarchy of theories determined by consistency strength, a finer-grained hierarchy in which theories higher up are capable of 'explaining', though not fully justifying, our commitment to theories lower down. One way to ascend Galvan's 'hierarchy of explanation' is to formalize soundness proofs: to this extent it often suffices to assume a full theory of truth for the theory T whose soundness is at stake. In this paper, we investigate the possibility of an extension of this method. Our ultimate goal will be to extend T not only with truth axioms, but with a combination of axioms for predicates for truth and necessity. We first consider two alternative strategies for providing possible worlds semantics for necessity as a predicate, one based on classical logic, the other on a supervaluationist interpretation of necessity. We will then formulate a deductive system of truth and necessity in classical logic that is sound with respect to the given (nonclassical) semantics.

1 Introduction

In recent decades, the techniques available for studying the functioning of the human brain have expanded enormously. Powerful new techniques such as neuroimaging (e.g., fMRI, PET, SPECT) allow neuroscientists to discover how mental states, cognition and emotion, are mapped on to neural substrata in the brain.

The more neuroscientists know about the functioning of the brain, the more they seem to abandon the assumption that there must be something like an Ego or a Self which is responsible for our actions. This leads to the suspicion that not we, but our brains, determine and control our doings; not we, but our brains do the thinking, planning, and acting. It is the brain's ability to perform all these functions that makes us human.

This is the message of the new naturalistic branch in philosophy called "neurophilosophy". Allied to neurophilosophy are social neuroscience, neuroe-

Edmund Runggaldier: Innsbruck University

DOI: 10.1515/9783110529494-020

conomics, neuroethics and even neurotheology. Adherents of neurotheology hope soon to be in a position to explain the origin of religious feelings and thought. "Neurotheology" has become a kind of keyword for neuroscientific progress in the field of religion. (See e.g. Alper (2008))

The message of neurophilosophy, that our Egos and Selves are ultimately illusions, is a great challenge for our traditional world view, for Christian anthropology and theology. This challenge has explicitly or at least implicitly naturalistic presuppositions: real entities can be described and explained in scientific terms. Should it turn out that entities cannot be described nor accounted for in a scientific context, they must be disregarded as unreal. The only means to achieve knowledge is science. Thus, if the personal Self or the human soul cannot be accounted for in scientific terms, it must be illusory.

For a naturalist, all features of the world are entirely caused by and realized in systems of micro-elements. Their properties and behaviour are sufficient to determine everything that happens in the macro-world. Every macro-phenomenon – human action included – has arisen through natural, micro-physical, causal processes. The existence of every macro-structure continues to depend causally on processes of this kind.

Today, we have a variety of reactions to this naturalistic challenge. Prominent are the answers given by the Christian Philosophers around the New Calvinists (Alwin Plantinga, William Alston). They are basically centered on a kind of retorsion: What happens if we apply the naturalistic stance to itself?

Sergio Galvan argues against the naturalistic presuppositions of the challenge on the basis of Gödels proofs. There is no complete language capable of describing all the real phenomena. A mere physicalistic or naturalistically restricted language won't do.

In this paper I will refer to a very popular strategy – at least in the German-speaking world – by resorting to the philosophy of Wittgenstein. Theologians, in particular, seem to be sympathetic to it. The advocates of a wittgensteinian strategy emphasize the peculiarities of the every-day language as used in practical and religious speech acts.

The strategy of adopting Wittgenstein's later philosophy of language (various "language-games" and "forms of life") undoubtedly has advantages. It is, however, in the long run, not convincing. As such it might lead to an isolation of Christian anthropology and to an unjustified immunization of theological propositions.

In the second part of the paper, I will refer to the strategy pursued in our Department, at Innsbruck University, of taking up the naturalistic challenge. It is a modest strategy which aims at elaborating the ontological presuppositions of the challenge. These are typical for a generalization of a scientific ontology.

Naturalists consider ontology to be a field of pure theoretical philosophy concentrating on the question of what the most fundamental constituents of reality are: What really exists are the "ultimates" of theoretical analysis. Thus, for them, ontology has to conform to the methods of science and exclude the kind of reasoning which lies within the practical domain of everyday life.

By contrast, we maintain – in agreement with the Aristotelian tradition – that ontology has to take into account the presuppositions of practical rationality as well, i.e. the subjective perspective and the first-person approach to reality. In the Aristotelian tradition, ontology has to account for both theoretical and practical rationality.

Thus, we concentrated on the every-day assumption that there is diachronic identity. The reasons for postulating that at least we, as human persons "endure" and are thus diachronically identical are based on our practical life. Decisive reasons for this view stem from our subjective experience of being intentional agents, planning our future and remembering our past.

It is obvious that natural science has to exclude subjectivity. Science is based on abstraction. It even abstracts from our experience of the passing time and thus considers time as a 4th dimension. The naturalistic ontology is 4-dimensional. We hold, however, that scientific 4-dimensionalism should be interpreted as a mere methodological abstraction. We are not forced by science to conclude that we, as personal agents, are 4-dimensional, extended not only in space but in time as well.

Our strategy is modest. We know that the reasons for questioning some of the presuppositions of naturalistic ontology are not scientific, but they result from taking into account the immediate experience of ourselves as acting.

But let us first see how wittgensteinians tend to respond to the mentioned naturalistic challenge stemming from neurophilosophy and its naturalistic assumptions.

2 The Wittgensteinian strategy

As mentioned in the introduction, the wittgensteinian way to respond to the naturalistic challenge is to point to the peculiarity of everyday¬ and religious language. Because of their special status, there cannot be a real opposition or contradiction between scientific claims on the one hand, and practical everyday and religious speech acts and beliefs on the other. This kind of reaction arises from the later Wittgenstein's philosophy of language, but is already present in his early philosophy of the *Tractatus*.

Both wittgensteinian positions yield the thesis that there cannot be real opposition between science and subjective and religious statements, but for different reasons. The first standpoint supports the view that subjective and religious speech acts merely appear to be statements, whereas in reality they are expressions of feelings and attitudes; the second standpoint, stemming from the Philosophical Investigations, stresses the view that everyday, first-person and religious speech acts belong to different language games and different forms of life.

According to the more radical position, we cannot make any cognitively relevant statements about our inner life and God. One has to distinguish between what can clearly be stated. i.e. scientifically described, and what can only be shown. Ultimately, it is best to keep silent where descriptive or scientific statements are not possible. Wittgenstein's dictum is well known: "Whereof one cannot speak, thereof one must be silent." (*Tractatus* 7)

This position seemingly neutralizes the challenges stemming from neuroscience and naturalism. There is no tension between scientific sentences and religious speech acts since the latter acts are not descriptive. With subjective and religious speech acts people do not make any cognitively relevant claims. The function of such claims is at most expressive. Religious language is significant insofar as it expresses a particular approach to reality, but religion cannot be denounced as wrong.

This wittgensteinian reaction and answer to the naturalistic challenge, that there is no Self and no God, has advantages, but amounts to an anti-cognitivist account of worldview and religious statements. There cannot be opposition between science and religion, but only because in a worldview (*Weltanschauung*) and in religion, ultimately, we do not describe anything.

In the long run, this position cannot be satisfactory. The main objection to any anti-cognitivist account of religious language relies on the fact that religious people do make claims to truth. Many religious believers not only believe their religious claims to be meaningful and true in their own context, but assume that they are true against other worldviews. If the anti-cognitivist standpoint were correct, a comparison between the meaningfulness and truth of worldviews and religious beliefs would be impossible.

The anti-cognitivist reaction to the naturalistic challenge is not convincing. Neither side, whether naturalistic or religious, is taken seriously. The naturalist must protest since she intends to make claims about what is the case and to separate it from mere fiction. She claims to have obtained knowledge that what is mistakenly taken to be real, i.e. the Self and the Divine, is not real. On the other side the faithful must protest, since for her at least some of her speech acts are cognitively relevant. By some of her speech acts she expresses more than feelings. If we take both seriously, the naturalist and the faithful, we have to take seriously

the possibility, at least in principle, of a cognitive tension between them. Many religious speech acts imply, for example, the proposition that God exists, but this is what the naturalist denies.

The more popular reaction to the naturalistic challenge is to resort to the later Wittgenstein. The supporters of this strategy hold a more moderate position by applying a pragmatic view of language based on Wittgenstein's theory of language games. In his *Investigations*, Wittgenstein proposes, as a way of establishing meaning, the analogy of games: just as each game has its own rules determining what can and cannot be done, so each language game has its own rules determining what is and is not meaningful.

Language, according to the later Wittgenstein, is ultimately an instrument for doing things. The different types of doings should not be confused or intermingled. What one does by using words depends on the context and differs accordingly from game to game.

The later Wittgenstein's understanding of language fits well into the tradition of pragmatic theories of meaning. To know the meaning of a word is to know how to use it, i.e. to have a command of its rules and to be able to follow them. The words used in different games can be the same, but they vary in their meaning, since the rules for their use vary. By using words, one describes not only states of affairs. One can ask, command, thank, greet, pray, etc.

Wittgenstein's account of language allows of a plurality of language games. Scientists have, as scientists, their own games, different from the games in which they speak of themselves as responsible agents or when they use religious language. We are thus entitled to classify first-person and religious speech acts as part of legitimate language games which are meaningful within their own context. These should not be confused with language games typical of science or of empirical descriptions. Thus, what people claim about their interior and about God in their language games cannot be challenged by scientific propositions.

The choice of a particular language game is not arbitrary. The language games are warranted by being embedded in a "form of life". (*Investigations* §19) Forms of life are the key to understanding the connection of language with everyday life. The speech within a given language game is embedded in the larger context of socio-cultural activity: "… the term 'language-game' is meant to bring into prominence the fact that the speaking of language is part of an activity, or of a form of life." (*Investigations* §23) Against the background of such an understanding of the interrelation between language and life, subjective and religious speech acts cannot be challenged by scientific statements, even less by their generalizations.

This reminds us of the neo-positivistic distinction between internal and external questions. The former are tackled within a frame of reference tied to clear criteria, whereas the latter refer to the status of the frames themselves (Carnap (1950a)). An internal question in an empirical framework is, for example, "Is there a white piece of paper on my desk?" The answer is to be found by looking. Or a question within the framework of natural numbers could be: "Is there a prime number smaller than 20?" The answer here is to be found by logical-mathematical method. Questions that relate to the frameworks as such, that are accordingly not asked inside a particular framework, as "Do numbers as such exist?, lack any cognitive relevance whatsoever. The choice of a framework depends solely on its practical role. If it is practically advantageous to use a particular frame of reference, one uses it.

Of course, logical positivists were convinced that the religious frameworks are not useful, but at least in theory they stressed – like the wittgensteinians – tolerance and the freedom of choice. If religious frameworks or language games should turn out not to be advantageous, they are rejected, but not for cognitive reasons. There is simply no objective reason for choosing one form of life or language game over another. "The form is given, the game is played, with its own rules, on its own field. The claims, assertions and practices within a language game or form of life cannot be fully understood from the context of another form of life, and there are no 'meta-criteria' standing above all forms of life that can decide between them." (Nicholson (1996), p. 629)

The concept of the wittgensteinian language games and the related notions suggest a form of linguistic and socio-cultural relativism. By developing the concept of meaning as use within a language game, which is played within a form of life, one tends to get rid of the relation of reference between an expression and an external reality. Against the background of Wittgenstein's later philosophy it does not make much sense to argue for or against the existence of God outside a religious form of life. Analogously arguments for or against the reality of the Self or the Ego are, outside an everyday language game, vacuous.

The theory of meaning based on the plurality of language games might be helpful in defending the relevant language games and as such in providing an answer to the naturalistic challenge: Religious language is different from language used to describe physical objects. Subjective personal claims on the reality of the Self and statements in the first person perspective cannot be criticized from the standpoint of empirical or natural science. Personal and religious claims cannot be challenged by science, not even by the new discoveries of neuroscience. There cannot be true opposition between claims of science on the one side and personal and religious speech acts on the other.

This wittgensteinian strategy of a defense against the naturalistic challenge, however, has many disadvantages stemming from its relativism and its failure to accommodate intuitions about objective truth. The strategy does not allow for arguments on the existence of the Self or God across the borders of different language games. The wittgensteinian theory of meaning reinforces the subjective conceptions of the Self and of religion as something insulated from and untouched by other language games. The corresponding insulation leads to a kind of immunity, protecting given claims from criticism.

On the philosophy of the language games there is a vast ongoing discussion. I mentioned broadly some of the intuitions of this philosophy supporting a very popular and seemingly tolerant reaction to the naturalistic challenge: Within some language games, claims to the reality of the Self and God are possible, but nonetheless this strategy, too, turns out to be inimical to truth claims, forestalling the possibility of cognitively relevant arguments across the games.

As mentioned in the introduction, at our Institute we perused a more modest response to the naturalistic challenge. We tried to clarify the presuppositions of the naturalistic challenge, arguing on the basis of the ontological presuppositions of practical everyday life. Even though subjectivity and the first person perspective have to be excluded from scientific methods, we maintain that they play an ontological role. The methodological abstraction in science does not imply that they lack ontological relevance. If we are right, the mere naturalistic 4-dimensional ontology has to be corrected. The subjective experience of consciously and intentionally acting (and by acting changing the world) conveys, so we believe, good reasons for assuming that we are 3-dimensional agents keeping our identity over time. We questioned the naturalistic ontological presupposition that there is no diachronic identity, its exclusion being one main reason for the mentioned naturalistic negation of the Ego or Self.

3 3-dimensional ontology (enduranatism)

Naturalistically-minded scientists treat time like a spatial dimension. For them everything that exists is spread out not only in space, but in time as well. Every real entity is extended in time too, i.e. composed of temporal stages/parts (see e.g. Quine (1960)). In addition to the 3 spatial dimensions of depth, length and height, naturalists consider time as a 4th dimension (Rea (2003)).

If everything is spread out in time, there cannot be diachronic identity. Thus, naturalists have to reject the everyday assumption that we, as agents, continue to exist in the “now” or in the ever-changing present. They have to reject the

belief that we remain the same even if we change. 4-dimensionalists have thus to reduce the assumption of persistence in time to a kind of continuity among adjacent temporal stages or parts. The continuity relation among temporal parts is weaker than diachronic identity and is neither reflexive nor transitive. It allows for differences of degree.

The 4-dimensional ontology, called "perdurantism", is mainly due, at least partially, to the successful use in science of, and the successful working with the 4-dimensional space-time system. The four-dimensional space-time system is useful for the representation of reality over time. However, its successful application does not imply that the world is four-dimensional.

According to the ontological position we adhere to, called "endurantism", real things, human persons included, are 3-dimensional. They "endure" and are thus called "endurers" going along with the "now". Being fully present at each moment of their existence they are neither extended in time nor composed of temporal parts.

4-dimensionalism conforms to conventionalistic empiricism. In this tradition it is commonplace to assume that individuals, living beings and human persons included, have to be constituted or "constructed". The constitution of individuals is conventional, its constraints being only pragmatic. Consistent conventionalism assumes this for the temporal dimension too. There is no "fact of the matter" about diachronic identity. In reality there are no individuals going with the "now".

As mentioned above, in science we need neither diachronic identity nor 3-dimensional "endurers". For an account of the presuppositions of practical philosophy, however, we need more. Everyday talk about ourselves, as responsible agents, supports the view that we need ontologies with "endurers" remaining identical throughout their existence. Who has done the deeds in the past? We or some temporal parts preceding our actual parts?

The decisive reasons for postulating "endurers" and thus diachronic identity are given by our subjective experience of being intentional agents planning, as said, our future and remembering our past. The presuppositions of practical rationality, agency, subjectivity, the first person perspective support endurantism. (Runggaldier (2006))

One special strategy for countering the naturalistic generalization of 4-dimensionalism, therefore, points out the disanalogies between space and time. As agents we are "prisoners" of time but not of space: we can choose where to live but not when. For an agent, time necessarily has a direction, whereas space does not.

Our uses of the indexicals "now" and "here" are disanalogous. An agent can now make choices about, deliberate about, and have intentions regarding her conduct later but not earlier than now. She can however make choices about,

deliberate about, and have intentions regarding her conduct in front of as well as behind her (Gale (2002), p. 70).

Our attitudes towards our past and future limitations are different. We prefer the painful experiences to be in the past and the pleasant ones in the future. Whereas we do not have any analogous spatial preference. It doesn't matter whether the pains and pleasures occur in any particular spatial direction from here.

Our emotional attitudes too reveal disanalogies between space and time. In our lives the fact that something has happened, is happening, or will happen is reason for how we feel. For instance, we feel deep grief on hearing that a beloved person has died. How could we account for anything like this deep grief by assuming within a 4-dimensional ontology that 'past', 'present', and 'future' events all have the same kind of reality (Cockburn, 1998, p. 84 f.)?

In everyday life we state events as reasons for feelings even if we do not know their objective dates. It often suffices to know that the relevant events are past, present or future. By using the indexical "now" we succeed in referring to the actual moment of time even if we do not know what time it is now. We do not need any objective time-references for that purpose. In a sense, by using indexicals, we even say more than by using their substitutes from the objective language of dates. It is impossible to say in a mere objective language what we mean when we say "Thank goodness that's over"!

My relief is not due to the fact that the event causing it takes place at a certain date, but that it is over. It is the "overness" or "pastness" of the event that I am thankful for. If I were not convinced that it is a fact that it is over I would not be relieved! The belief that something has happened in the past or will happen in the future provides us with reasons not only for acting in a certain way, but for feeling certain emotions too. Our indexical beliefs and our experiences of acting and feeling favor and support the thesis that at least we are "endurers" and are thus diachronically identical with ourselves.

In real life we do not constitute/construct agents out of temporal stages, but we presuppose that they go on, always existing in the "now". We form judgements presupposing diachronic identity, not inferring them from the alleged distribution of stages or temporal parts. In our attempt at defending 3-dimensionalism, however, we neither ignore nor negate that in order to yield objective results, natural science has to exclude subjectivity, indexicality and the first-person perspective. But, as said, this should be interpreted as a methodologically convenient assumption. It does not force us to conclude that there is neither subjectivity nor indexicality and that there are no 3-dimensional "endurers".

4 Concluding remarks

If the achievements of neuroscience in understanding how the brain works and in identifying the substrata of mental states are interpreted naturalistically they challenge the Christian anthropology and theology. Naturalists, in fact, stick to the thesis that neuroscience has shown that the assumption of an Ego or a Self or God is illusory: If our personal Self or the human soul cannot be accounted for in scientific terms it must be an illusion!

One very popular strategy for responding to this naturalistic challenge is wittgensteinian. However this response tends to block any significant dialogue between neuroscience on the one hand and philosophical anthropology and theology on the other. In the long run it is not satisfactory to thoroughly separate the everyday and religious language games from those of science.

In our Department we hold that the dialogue between natural science and anthropology and theology is not obsolete: it is possible to argue both for and against the theses of neurophilosophy. Understanding how the brain works does in any case not suffice as an adequate account of mental phenomena. Even if successful, neurotheology, for example, would at best explain how the brain develops religious feelings and thoughts. It surely is not sufficient for negating the existence of God.

The exclusion of diachronic identity is one main reason for the naturalistic negation of the Ego or Self. We thus concentrated on the thesis that human agents are 3-dimensional "endurers". Our modest strategy in reacting to the naturalistic challenge consisted basically in negating its ontological presupposition that there is no diachronic identity.

The 4-dimensional space-time system is undoubtedly very useful for scientific purposes and for the representation of reality, but its successful use does not imply that everything is four-dimensional. The decisive reasons against the generalization of 4-dimensionalism and for postulating 3-dimenisonal "endurers" and thus diachronic identity stem from our experience of acting, from practical rationality, subjectivity, and indexicality. If so, we are not forced on ontological grounds to adhere to a purely naturalistic worldview. The findings of neuroscience do not force us to believe that our Selves or Egos are illusions and that there is no God.

Gerhard Schurz and Ernest Adams (1926–2009)

Measure-Entailment and Support in the Logic of Approximate Generalizations

Abstract: Adams' logic of conditional epistemic probabilities (or "degrees of belief") is well-known. However, there is also a logic of statistical probabilities expressing facts about the real world. In this paper these probabilities are called "degrees of truth" of open formulas that are considered as *approximate generalizations*. Based on Adams (1974), this paper presents a precise elaboration of the logic of unconditional and conditional degrees of truths, based on a measure-theoretic notion of entailment and support, including a restricted completeness proof and applications to the theory of orderings, balanced structures and the notion of quasi-classical reasoning.

1 Introduction

In Adams (1974, 1975, 1977, 1986) the logic of unconditional and conditional probabilities has been developed. These probabilities are *epistemic* probabilities, also called "degrees of belief". However, there is also a logic of unconditional and conditional *statistical* probabilities. Since these probabilities express facts about the real world, we call them in this paper *degrees of truths* of open formulas

Ernest Adams is well-known for his pioneering work on the logic of conditional epistemic probabilities. Less known is his work on the logic of statistical probabilities, to which Adams refers as "degrees of truth" of open formulas that are considered as approximate generalizations. This paper emerged from my work with Ernest W. Adams on the logic of approximate generalizations in the years 1995–2000. The paper was completed in 2001 and was submitted as a joint paper to the JPL, when Ernest Adam's state of health was severely impaired and the publication of the paper was delayed. Ernest Adams died in 2009. I dedicate the publication of this paper to Ernie whose pioneering work on the probabilistic semantics for conditionals had a sustainable influence on my own philosophical development.

Gerhard Schurz: Chair of Theoretical Philosophy, University of Düsseldorf, Universitätsstrasse 1, Gebäude 24.52, D-40225 Düsseldorf, Germany E-Mail: schurz@phil.uni-duesseldorf.de
Ernest Adams (1926–2009): Formerly, Department of Philosophy, University of California at Berkeley, Berkeley, California 94720, USA

DOI: 10.1515/9783110529494-021

that are considered as approximate generalizations. Basic ideas concerning this logic have been laid down in Adams (1974). The following paper presents a precise elaboration of the logic of unconditional and conditional degrees of truths, based on a measure-theoretic notion of entailment and support, including a restricted completeness proof and applications to the the theory of orderings, balanced structures and the notion of quasi-classical reasoning.

The open formula $xRy \to xSy$ (with $\to$ for material implication) may not be true for all pairs $< x, y >$, but only for most of them, for instance, when R is the 'as least as great as' relation among numbers of a large class, and S is the 'greater than' relation. In this case $xRy \to xSy$ is regarded as an *absolute approximate generalization* (*a.a.g.*). Adams (1974) and Carlstrom (1975) have developed a logic of these expressions that reflects their approximate character by measuring their *degrees of truth* $t(-)$ by the proportion of values of their free variables that satisfy them. In many cases, however, it is more plausible to consider a generalization of the form "if x has relation R to y then it has relation S to it" as a *conditional approximate generalization* (*c.a.g.*), which is formalized with help of a non-material conditional operator $\Rightarrow$, as $xRy \Rightarrow ySx$. Its degree of truth is measured by the proportion of values of their free variables which satisfy their consequent out of the values that satisfy their antecedent. For example, if we interpret 'xRy' as 'x is related to y' in the domain of people, and 'xSy' as 'x is the son of y', then $t(xRy \to xSy)$ is high because $xRy \to xSy$ is true when xRy is false and xRy is highly false, while $t(xRy \Rightarrow xSy)$ is low because most pairs of related people are not related as parent and son. For finite domains, the degrees of truth of a.a.g.s and c.a.g.s are their absolute or conditional frequencies, respectively, while for infinite domains, appropriate measure-theoretical generalizations are needed.

Focusing on degrees of truth rather than on truth *simpliciter* requires replacing the standard concept of entailment by a more general concept of *measure-entailment* (m-entailment), roughly as follows. A.g.s $\varphi_1, \ldots, \varphi_n$, either absolute or conditional, m-entail another a.g., φ, which may also be either absolute or conditional, if the degree of truth of φ can be guaranteed to be arbitrarily close to 1 in a model, by requiring the degrees of truth of $\varphi_1, \ldots, \varphi_n$ to be 'sufficiently' close to 1 in the model. Most of the results stated in Adams (1974) and Carlstrom (1975) pertain to m-entailment in the case of absolute a.g.s. This paper concerns extensions of these results, in particular concerning a rather singular phenomenon that seems to have no parallel in classical logic.

An inference that will be considered at greater length in subsequent sections has only one premise, namely the disjunction $xRy \vee yRx$, which is highly true, for instance when xRy expresses the fact that x is greater than y among numbers.

Now, it is obvious that this premise does not m-entail xRy, that for most x and y, x has the relation to y, since that is not highly true of the greater-than relation. However, the curious fact is that given that $xRy \vee yRx$ is highly true, xRy cannot be highly untrue. In fact, if $xRy \vee yRz$ is 99% true (true for 99% of the pairs of values of x and y), then it follows from simple laws of the algebra of degrees of truth that xRy must be at least 49.5% true. While $xRy \vee yRx$ does not entail xRy, it 'supports' it by guaranteeing that it has a non-negligible degree of truth. This is interesting for two reasons. First, measure-support (in short: support) differs from both deductive and inductive support and has not been considered heretofore, and does not arise at all in most of the calculi of approximate or fuzzy truth in the current literature. And second, considering this phenomenon brings out a connection between the degrees of truth of absolute and conditional a.g.s that is important in applications. The latter will be discussed in subsequent sections, but first we will lay down the foundations of m-entailment and support relations among absolute and conditional a.g.s.

2 Measure-entailment among absolute a.g.s

Syntactically, an a.a.g. may be any formula φ of the first order predicate language. What makes φ an a.a.g. is not is syntactical structure, but its semantical evaluation. For reasons of simplicity, we exclude function symbols (in completeness theorems, we must also exclude identity). We assume a standard concept of models $M =< D, I >$ for such a language, where D is the domain and the interpretation function I assigns appropriate values to the nonlogical symbols of the language. Given a model M and a formula $\varphi = \varphi(x_1, \ldots, x_n)$ with free variables $x_1, \ldots, x_n$, its *degree of truth in* M, $t_M(\varphi)$, is the proportion of n-tuples $< d_1, \ldots, d_n >$ in D that satisfy φ in M. If φ is a sentence without free variables then $t_M(\varphi)$ is defined as 1 or 0, according as φ is true or false in M. In most of the cases that we will consider D will be finite and $t_M(\varphi)$ is defined, but if D is infinite then $t_M(\varphi)$ must be generalized by introducing a σ-additive measure, μ, over a Borel class of subsets of D, where $t_{M,\mu}(\varphi)$ is defined as the measure of the class of n-tuples $< x_1, \ldots, x_n >$ in the μ^n (the n-th power of μ) that satisfy φ in M. For convenience we will omit the subscripts of 't'; thus when we speak of a given (or of every, or of some) degree-of-truth function $t(-)$ we implicitly refer to a given (or to every, or to some, respectively) pair $< M, \mu >$. For given $t(-)$, $f(-) =_{df} 1 - t(-)$ is the corresponding degree-of-falsity function.

Degrees of truth satisfy the axioms of probability: (1) $t(\varphi) \geq 0$, (2) $t(\varphi \upsilon \neg\varphi) = 1$, and (3) if φ and ψ are logically inconsistent then $t(\varphi \vee \psi) = t(\varphi) + t(\psi)$ (among

other things, this entails and (4) $t(\neg\varphi) = 1 - t(\varphi)$ and (5) if φ logically implies ψ, then $t(\varphi) \geq t(\psi)$).[1] *However*, and this must be stressed, degrees of truth also satisfy certain laws that are not satisfied by probabilities in general. First, if φ and φ' are *alphabetic variants* of one another (e.g., xRy and yRx), then $t(\varphi) = t(\varphi')$. And, if φ and ψ *have no free variables in common* then $t(\varphi \wedge \psi) = t(\varphi) \cdot t(\psi)$. In effect, these are the properties of *symmetry* and *independence*. We may add that de Finetti's Theorem implies that any symmetric probability function over the language is a mixture of independent and symmetric probability functions. This allows us to regard symmetric probability functions on a language as mixtures of degree of truth functions on it, and we may also say that degrees of truth are 'factual' or 'statistical' probabilities, while 'real' probabilities as characterized by axioms (1)-(3) are 'epistemic' probabilities and depend on what one knows about the facts. While recent work on 1st order probability logics was almost exclusively about epistemic probabilities[2], this paper concerns 1st order probability logics for statistical probabilities, i.e., for degrees of truth.

A finite class of formulas $\varphi_1, \ldots, \varphi_n$ *m(easure)*-entails another formula, ψ, abbreviated as $\varphi_1, \ldots, \varphi_n \Vdash_m \psi$, iff for all $\epsilon > 0$ there exists $\delta > 0$ such that for all degree-of-truth functions t, if $t(\varphi_i) \geq 1 - \delta$ for $i = 1, \ldots, n$, then $t(\psi) \geq 1 - \epsilon$. From now on we assume that ϵ and δ range over small but non-zero real numbers. Certain key properties of m-entailment among a.a.g.s are demonstrated in Adams (1974), including the calculus **M** for m-derivability ($\vdash_m$). Some terminology: $\Vdash$ stands for classical first order consequence in the sense that $\varphi_1, \ldots, \varphi_n \Vdash \psi$ iff for all models $M =< D, I >$ and valuation functions v (which assigns elements of D to the variables of the language) it holds that if(M, v) satisfies all of the φ_i, then (M, v) satisfies ψ. $\vdash$ is the corresponding notion of first order derivability (complete for $\Vdash$). An *alphabetic variant* $var(\varphi)$ of formula φ results from φ by replacing free variables by other distinct free variables; so '*var*' may be considered as a permutation function on the (free) variables of the language. A *variable-condensement* $c_\varphi(\psi)$ of formula ψ with respect to formula φ results from ψ by replacing some free variables of ψ which are not free in φ by distinct free variables which are free in φ but not free in ψ. For example, Qxy is a condensement of Qxz w.r.t. Rxy, but Qxx is neither an alphabetic variant of Qxz nor a condensement of Qxz w.r.t. Rxy.

1 Axiom (3) generalizes to σ-additivity in the case of infinite sequences of formulas $\varphi_1, \varphi_2, \ldots$.
2 Cf. Hawthorne's [1998] development of a 1st order conditional probabilistic logic, and Hailperin's [2000] system of 1st order probability logic.

Calculus M: $\varphi_1, \ldots, \varphi_n \vdash_m \psi$ iff ψ is derivable from $\varphi_1, \ldots, \varphi_n$ by the following three rules of inference (AV alphabetic variant, LC logical consequence, DC disjunctive condensement):

(AV) $\varphi \vdash_m var(\varphi)$, for every alphabetic variant $var(\varphi)$
(LC) If $\varphi_1, \ldots, \varphi_n \vdash \psi$, then $\varphi_1, \ldots, \varphi_n \vdash_m \psi$
(DC) $\varphi \vee \psi \vdash_m \varphi \vee c_\varphi(\psi)$, for every condensement $c_\varphi(\psi)$

M-entailment expresses the *infinitesimal* behaviour of degrees of truth (i.e., what happens with t of the conclusion if t of the premises goes to 1). We are also interested in their *non-infinitesimal* behaviour, which is expressed in terms of lower bounds of the conclusion's degree of truth – or equivalently, in terms of upper bounds of the conclusion's degrees of falsity – in dependence of the premises' degrees of truth (or falsity, respectively). We may also say that non-infinitesimal behaviour concerns the *degree of preservation of degrees of truth*. Rule (AV) strictly preserves the degree of truth (and falsity). For inferences of classical logic (rule (LC)) it is a well-known fact that the conclusions' degree of falsity, $f(\psi)$ is smaller-equal than the sum of the premises' degrees of falsity (cf. Adams 1965, Suppes 1965). For rule (DC), the degree of f-preservation is bad: it holds (for all degrees-of-truth functions) that $f(\varphi \vee c_\varphi(\psi)) \leq \sqrt{f(\varphi \vee \psi)}$. For example, if $t(Rxy \vee Qxz) = 0.9$, then $t(Rxy \vee Qxy) \geq 1 - \sqrt{0.1} = 0.66$. This follows from an appropriate generalization of the measure-theoretic fact that $t(Fxy \wedge Gxz) \geq (t(Fxy \wedge Gxy)^2$, via $f(\varphi \wedge \psi) = t(\neg\varphi \wedge \neg\psi) \geq (t(\neg\varphi \wedge \neg c_\varphi(\psi))^2$ (cf. Adams 1974, p. 6; the fact results from an application of the law of projection and Tschebyscheff's inequality).

In m-entailment among *absolute* a.g.s, the free variables may be implicitly understood as being quantified by an invisible *most-quantifier* (M) which was explicitly introduced in Adams [1974]. It is therefore useful to compare m-entailment with *universal* classical entailment $\models_u$, where free variables are implicitly understood to be quantified by an invisible strict universal quantifier ($\forall$), defined as follows: $\varphi_1, \ldots, \varphi_n \models_u \psi$ iff for all models M, if the φ_i's are satisfied in M by all valuations of variables, then ψ is satisfied by all valuations of variables. Neither $\models_u$ nor $\models_m$ satisfies the standard structural rules of classical consequence, because both entailment relations refer to implicitly quantified formulas. As an example, consider the invalidity of Case Distinction: $Fx \models_u Fx \vee \neg Fy$ and $\neg Fx \models_u Fx \vee \neg Fy$, but $Fx \vee \neg Fx \not\models_u Fx \vee \neg Fy$. Likewise, $Fx \models_m Fx \vee \neg Fy$ and $\neg Fx \models_m Fx \vee \neg Fy$, but $Fx \vee \neg Fx \not\models_m Fx \vee \neg Fy$. On similar reasons, deduction theorem and contraposition are invalid for both entailment relations. However, Cut, Monotony and Modus Ponens are valid for both of them.

Both entailment relations are strictly stronger than classical entailment $\Vdash$ (e.g., $Fx \Vdash_u (\Vdash_m) Fy$, but $Fx \not\Vdash Fy$). Moreover, $\Vdash_u$ is strictly stronger than $\Vdash_m$: while

$\vdash_u$ satisfies the unrestricted law of substitution (Subst): $\varphi \vdash_u \varphi'$, where φ' results from φ by substituting any terms for variables in φ, $\vdash_m$ satisfies only the weaker rules (AV) and (DC), which exclude substitutions by constants and allow mergings of variables only in the special case of (DC). For example, $Rxy \vdash_u Rzx$, $Rxy \vdash_u Rza$ and $Rxy \vdash_u Rxx$; while $Rxy \vdash_m Rzx$, $Rxy \nvdash_m Rza$, $Rxy \nvdash_m Rxx$.[3]

A *standard derivation* of ψ from $\varphi_1, \ldots, \varphi_n$ is a derivation which consists, first, in a number of applications of rule (AV) to the premises, second, in a series of applications of rule (LC), arriving at a disjunction, to which, third, a series of applications of rule (DC) is applied until one arrives at a disjunction of the form $\psi \vee \ldots \vee \psi$ from which ψ follows by one additional (LC)-step of $\vee$-contraction. The following theorem generalizes the results of Adams (1974, theorem 1) in two respects: theorem 2.1.(6) demonstrates how a standard m-derivation can be used to calculate a general upper bound for the conclusion's degree of falsity from given degrees of falsity of the premises, and theorem 2.1.(5) shows how this fact can be used as a *non-infinitesimal* probability semantics for m-entailment in the sense of Schurz (1998). These extensions are not only important for practically reliable reasoning with calculus **M** (cf. Schurz 1997); they will also be useful for the notion of support.

Theorem 2.1. *Let $\varphi_1, \ldots, \varphi_n, \psi$ be formulas without identity, where $x_1, \ldots, x_m$ are the variables free in ψ. Then the following conditions are equivalent:*[4]

(1) $\varphi_1, \ldots, \varphi_n \vdash_m \psi$

(2) $\varphi_1, \ldots, \varphi_n \vdash_m \psi$

(3) It is not the case that there exist disjoint sets $D, D_1, \ldots, D_m$ where D is denumerable and $D_1, \ldots, D_m$ are denumerably infinite, and a model M with domain $D \cup D_1 \cup \ldots \cup D_m$ such that (i) $\varphi_1, \ldots, \varphi_n$ are satisfied in M under all valuations which assign distinct values in $D \cup D_1 \cup \ldots \cup D_m$ to distinct free variables, and (ii) ψ is false in M under all valuations v such that $v(x_i) \in D_i$ for $1 \leq i \leq m$.

(4) It is not the case that for all δ there exists a degree-of-truth-function t such that $t(\varphi_i) \geq 1 - \delta$ for $1 \leq i \leq n$, but $t(\psi) \leq 1 - m^{-m}$.

(5) There exists natural numbers $k_1, \ldots, k_n, r$ such that for all degree-of-truth functions, $f(\psi) \leq (\Sigma\{k_i \cdot f(\varphi_i) : 1 \leq i \leq n\})^{(1/2^r)}$.

---

3 A further difference consists in the fact that while with $(\forall)$ of standard first order logic we may strictly generalize over particular variables, with (M) we may only most-generalize over all free variables simultaneously. An extension to expressions such as $(Mx)(My)Rxy$ ('most x are such that most y bear relation R to it') is a task for the future.

4 Carlstrom (1975) shows that m-entailment is compact, i.e. $1 \Leftrightarrow 2$ of theorem 2.1 holds also for infinitely many premises; the restriction to the finite case is needed for theorem 2.1.(5,6).

(6) *There exists a standard-derivation of* ψ *from* $\varphi_1, \ldots, \varphi_n$ *such that for all degree-of-truth functions,* $f(\psi) \leq (\Sigma\{k_i \cdot f(\varphi_i) : 1 \leq i \leq n\})^{(1/2^r)}$, *where* k_i *is the number of distinct alphabetic variants of* φ_i *introduced by (AV), and* r *is the number of (DC)-steps.*

Proof. We prove the equivalence of (1)-(6) by establishing the circle of implications (1) ⇒ (4) ⇒ (3) ⇒ (6) ⇒ (5) ⇒ (1), and the implications (2) ⇒ (1), (6) ⇒ (2). The implications (1) ⇒ (4) ⇒ (3) and (2) ⇒ (1) are proved as in Adams (1974, pp. 8f). The implications (6) ⇒ (5) ⇒ (1) and (6) ⇒ (2) are obvious. It remains to prove (3) ⇒ (6), by establishing ¬(6) ⇒ ¬(3) (similar to the proof of ¬ (2) ⇒ ¬ (3) in Adams 1974) as follows.

Assume no standard derivation of ψ from $\varphi_1, \ldots, \varphi_n$ exists. For all $1 \leq i \leq m$, let $V_i = \{x_{i,1}, x_{i,2}, \ldots\}$ be pairwise disjoint denumerably infinite sets of variables, with $x_{i,1} := x_i$, and let VAR be the set of alphabetic variants of premises $\varphi_i (1 \leq i \leq n)$ in the set of variables $V_1 \cap \ldots \cap V_m$. Let CONC be the set of conclusion-variants of the form $\psi(x_{1,k_1}, x_{2,k_2}, \ldots, x_{m,k_m})$, for arbitrary $k_1, \ldots, k_m$ such that $x_{i,k_i} \in V_i$; in other words, every variable x_i of ψ gets replaced by some variable in V_i. From every disjunction of elements of CONC the disjunction $\psi \vee \ldots \vee \psi$ follows by iterated applications of rule (DC) (part 3 of a standard derivation), while every finite subset of VAR follows from the premises by applications of (AV). Since there exists no standard-derivation, and by compactness of classical first order logic, it follows that VAR and the set of all negations of formulas in CONC must be logically consistent. As described in Adams (1974, p. 9), this fact implies the negation of (3); i.e., there exists such a model whose existence is excluded in (3).

So, by contraposition, (3) implies the existence of a standard-derivation of $\varphi_1, \ldots, \varphi_n \vdash \psi$. From the latter fact, the upper bound of the conclusion's degree-of-falsity as specified in (6) follows easily from the non-infinitesimal bound behaviour of the three rules as explained above.

qed

If identity is included, the proof breaks down since neither the model **M** can be assumed to be infinite nor the valuation function v can be assumed to be injective. The rules of **M** are also valid also for formulas with identity[5], but **M** is then no longer complete. For example, let ORDINF be the first order formula asserting that R is an irreflexive, asymmetric and transitive relation without greatest element. Then ORDINF is only true in infinite domains. So ORDINF $\Vdash_m \neg x = y$ is m-valid,

5 In particular, all classical entailments among identity formulas are m-valid. For example, $\Vdash_m x = x$ because $t(x = x) = 1$; $x = y, Rxy \Vdash_m Rxx$ because $t(x = y) = 1$ implies that the size of the domain is 1; $Rxy, \neg Rxx \Vdash_m \neg x = y$ because $t(Rxy) = 1, t(Rxx) = 0$ can only hold in infinite domains.

because the set of tuples $\langle x, y\rangle$ satisfying $\neg x = y$ has degree of truth 1 if and only if the domain is infinite, but this inference is not derivable in **M**.[6]

3 Support among absolute a.g.s

Definition 3.1.

(1) $\varphi_1, \ldots, \varphi_n$ *support* ψ *at level* α *(where* $\alpha \geq 0$*) iff for all* $\epsilon > 0$ *there exists* $\delta > 0$ *such that for all degree-of-truth-functions t, if* $t(\varphi_i) \geq 1 - \delta$ *for* $i = 1, \ldots n$*, then* $t(\psi) \geq \alpha - \epsilon$.

(2) $\varphi_1, \ldots, \varphi_n$ *support* ψ *if and only if they support* ψ *at some positive level.*

Note that if formula set Δ supports φ at level α, then Δ supports φ at every level β below α. Theorem 3.1 below characterizes support in analogy to theorem 2.1. Theorem 3.1.(2,3) show how the concept of support can be reduced to a condition about m-entailment, namely m-inconsistency. The concepts of m-insatisfiability and m-inconsistency are defined as one defines them in classical logic; see theorem 3.1.(2,3). By completeness (theorem 2.1), both notions coincide. Moreover, the argument for $\neg(2) \Rightarrow \neg$ (1), in the proof of theorem 3.1, tells us that a finite formula set Γ is m-satisfiable iff for every ϵ there exists a degree-of-truth-function t such that for all $\gamma \in \Gamma$, $t(\gamma) \geq 1 - \epsilon$. Recall that the fact that φ and ψ are m-inconsistent, i.e. that φ, $\neg\psi$ m-entail $\bot$, does not imply that φ m-entails ψ, because m-entailment does not satisfy the rule of Reductio ad Absurdum. However, that $\varphi, \neg\psi$ m-entail $\bot$ implies that φ supports ψ. Theorem 3.1.(3) gives a syntactic characterization: a derivation of 'φ supports ψ' consists in a derivation of '$\varphi, \psi \vdash_m \bot$'. Theorem 3.1.(5) tells us that a standard-derivation for support consist just in a series of (AV)-steps follows by (LC)-steps; (DC)-steps are never needed, because $\bot$ contains no variables. We call such a derivation an AV-LC-derivation.

Theorem 3.1. *The following conditions are equivalent:*

(1) $\varphi_1, \ldots, \varphi_n$ *support* ψ.

(2) $\varphi_1, \ldots, \varphi_n, \neg\psi$ *are m-insatisfiable; i.e.,* $\varphi_1, \ldots, \varphi_n, \neg\psi \Vdash_m \bot$.

(3) $\varphi_1, \ldots, \varphi_n, \neg\psi$ *are m-inconsistent; i.e.* $\varphi_1, \ldots, \varphi_n, \neg\psi \vdash_m \bot$.

(4) Either $\varphi_1, \ldots, \varphi_n \Vdash_m \psi$*, or* $\varphi_1, \ldots, \varphi_n$ *support* ψ *at most at level* $1 - m^{-m}$*, where m is the number of free variables in* ψ.

6 Carlstrom (1975) has proved completeness for formulas including identity but excluding quantifiers.

(5) *There exist numbers* $k_1, \ldots, k_n, r$ *such that for all degree-of-truth-functions* t, $t(\psi) \geq (1/r) - (\Sigma\{k_i \cdot f(\varphi_i) : 1 \leq i \leq n\})/r$.

(6) *There exists a AV-LC-derivation of* $\varphi_1, \ldots, \varphi_n, \neg\psi \vdash_m \bot$ *such that (i)* $\varphi_1, \ldots, \varphi_n$, *supports* ψ *at level* $(1/r)$, *where* r *is the number of variants of* $\neg\psi$ *used in this derivation, and (ii) for all degree-of-truth-functions* t, $t(\psi) \geq (1/r) - (\Sigma\{k_i \cdot f(\varphi_i) : 1 \leq i \leq n\})/r$, *where* k_i *is the number of variants of* φ_i *used in this derivation.*

Proof. Using theorem 2.1 we prove the equivalence of (1)-(6) by proving the two circle of implications (2) ⇒ (6) ⇒ (5) ⇒ (1) ⇒ (2), (1) ⇒ (4) ⇒ (1), and the two implications (3) ⇒ (2), (6) ⇒ (3).

(2) ⇒ (6): By (2) and theorem 2.1 (2 ⇔ 6), there exists a standard derivation of $\varphi_1, \ldots, \varphi_n, \neg\psi \vdash_m \bot$. Since the conclusion $\bot$ has no free variables, no (DC)-step can occur in this derivation; so it must be an AV-LC-derivation. By theorem 2.1.(2 ⇔ 6) it follows that for all degree-of-truth-functions t, $f(\bot) = 1 \leq \Sigma\{k_i \cdot f(\varphi_i) : 1 \leq i \leq n\} + r \cdot f(\neg\psi)$, where k_i and r are explained as in theorem 3.1.(6). By simple algebraic transformation, condition (ii) of theorem 3.1.(6) follows. From (ii) and the definition of support we directly obtain $1/r$ as a lower bound of the level of support, hence condition (i).

(6) ⇒ (5) ⇒ (1) is obvious. To establish (1) ⇒ (2) we prove ¬(2) ⇒ ¬(1): That $\varphi_1, \ldots, \varphi_n, \neg\psi$ does not m-entail $\bot$ implies by definition that there exists an ϵ such that for all δ there exists a degree-of-truth function t such that $t(\varphi_1), \ldots, t(\varphi_n) \geq 1 - \delta$, $t(\neg\psi) \geq 1 - \delta$ and thus $t(\psi) < \delta$, and $t(\bot) \leq 1 - \epsilon$ (which is trivially satisfied for arbitrary $\epsilon \leq 1$). But this implies, by definition of support, that there exists no positive level α at which $\varphi_1, \ldots, \varphi_n$ could support ψ.

(1) ⇒ (4) follows directly from theorem 2.1.(4), and (4) ⇒ (1) holds since both disjuncts of (4) imply (1). (3) ⇒ (2) follows from theorem 2.1, and (6) ⇒ (3) is obvious. qed

That in standard-derivations of support-relations (DC)-steps are never needed (th. 3.1.(6)) may sound surprising: for example, $Rxy \vee Qxz$ m-entails $Rxy \vee Qxy$ by rule (DC), and hence supports it. But that $Rxy \vee Qxz$ supports $Rxy \vee Qxy$ can be derived without using (DC) by the AV-LC-derivation $\{Rxy \vee Qxz, \neg(Rxy \vee Qxy)\} \vdash_{AV} \{Rxy \vee Qxz, Rxz \vee Qxy, \neg(Rxy \vee Qxy), \neg(Rxz \vee Qxz)\} \vdash_{LC} \bot$. (Here, $\Gamma \vdash \Delta$ means that $\Gamma \vdash \delta$ holds for every $\delta \in \Delta$, and $\vdash_{AV}$ means derivation by using AV-steps only.) The AV-LC-derivation of $\bot$ gives us a good (but not necessarily best) level of support according to theorem 3.1.(6), which here is 1/2. It gives us also a non-infinitesimal lower bound of the degree-of-truth for the supported conclusion, namely $1/2 - f(Rxy \vee Qxz)$. Theorem 3.1.(4) states a general upper bound for the support-level, based on theorem 2.1.(4), which is independent from the way in which $\bot$ was

derived; it is usually higher than the support-level calculated from theorem 3.2.(5), in our case it is $1 - 2^{-2} = 3/4$.

The next theorem, 3.2, collects some useful facts. Theorem 3.2.(1) etablishes m-entailment as the limiting kind of support with level 1 (hence m-entailment implies support). Theorem 3.1.(3) implies the important theorem 3.2.(2) which tells us that for conclusions with just one free variable the notions of m-entailment and support collapse. In other words, considerations of support have only significance for multi-variable formulas. The other parts of theorem 3.2 will be used in later sections. Theorem 3.1.(6) generalizes the fact which was recognized in the introduction: that $xRy \vee yRx$ supports xRy at level 1/2. Theorem 3.2.(7) explains a special case where the support-relation is transitive.

Theorem 3.2.

(1) $\varphi_1, \ldots, \varphi_n$ *support* ψ *at level 1 iff* $\varphi_1, \ldots, \varphi_n$ *m-entail* ψ.

(2) If ψ *has at most one free variable, then* $\varphi_1, \ldots, \varphi_n$ *support* ψ *iff* $\varphi_1, \ldots, \varphi_n$ *m-entail* ψ.

(3) If each of $\varphi_1, \ldots, \varphi_n$ *m-entails each of* $\psi_1, \ldots, \psi_k$, *respectively, then* $\varphi_1 \vee \ldots \vee \varphi_n$ *supports* $\psi_1 \vee \ldots \vee \psi_k$.

(4) If $\varphi_1, \ldots, \varphi_n$ *m-entail* ψ *and* ψ *supports* π, *then* $\varphi_1, \ldots, \varphi_n$ *support* π.

(5) If $\varphi_1, \ldots, \varphi_n$ *m-entail* $\psi \rightarrow \pi$ *and support* ψ, *then* $\varphi_1, \ldots, \varphi_n$ *support* π.

(6) If $\mathrm{var}_1(\varphi), \ldots, \mathrm{var}_n(\varphi)$ *are alphabetic variants of* φ, *then* $\mathrm{var}_1(\varphi) \vee \ldots \vee \mathrm{var}_n(\varphi)$ *support* φ *at level* $1/n$.

(7) If $\mathrm{var}_1(\varphi), \ldots, \mathrm{var}_n(\varphi)$ *are alphabetic variants of* φ, *and* $\psi_1, \ldots, \psi_k$ *support* $\mathrm{var}_1(\varphi) \vee \ldots \vee \mathrm{var}_n(\varphi)$ *at level* α, *then* $\psi_1, \ldots, \psi_k$ *support* φ *at level* α/n.

Proof. (1) follows from definition, and (2) is a direct consequence of theorem 3.1.(4) with $n = 1$. *Concerning (3)*: suppose for reductio that φ_i m-entails ψ_j for each $1 \leq i \leq n$ and $1 \leq j \leq k$, but $\psi_1 \vee \ldots \vee \psi_k$ is not supported by $\varphi_1 \vee \ldots \vee \varphi_n$. Then $\neg(\psi_1 \vee \ldots \vee \psi_k) \wedge (\varphi_1 \vee \ldots \vee \varphi_n)$ is m-consistent by theorem 3.1.(2), whence $[\varphi_1 \wedge \neg(\psi_1 \vee \ldots \vee \psi_k)] \vee \ldots \vee [\varphi_n \wedge \neg(\psi_1 \vee \ldots \vee \psi_k)]$ is m-consistent, and therefore $[\varphi_1 \wedge \neg\psi_1] \wedge \ldots \vee [\varphi_n \wedge \neg\psi_k]$ is m-consistent. But if φ_i m-entails ψ_j (for each i and j) then $\varphi_i \wedge \neg\psi_j$ must be m-inconsistent, and so must be $[\varphi_1 \wedge \neg\psi_1] \vee \ldots \vee [\varphi_n \wedge \neg\psi_k]$. (4) is clear by the definitions of m-entailment and support. *Concerning (5)*: Assuming the if-part, we have (a) $\varphi_1, \ldots, \varphi_n \Vdash_m \psi \rightarrow \pi$ and, by theorem 3.1(2), (b) $\varphi_1, \ldots, \varphi_n, \neg\psi \Vdash_m \bot$. From (a) we obtain (c) $\varphi_1, \ldots, \varphi_n, \neg\pi \Vdash_m \neg\psi$ (by LC-steps). (c) and (b) give us (d) $\varphi_1, \ldots, \varphi_n, \neg\pi \Vdash_m \bot$ (again by LC-steps), which implies by theorem 3.1.(2) that $\varphi_1, \ldots, \varphi_k$ support π. (6) is clear from theorem 3.1.(6), because we need n variants of $\neg\varphi$ to derive a contradiction with $\mathrm{var}_1(\varphi) \vee \ldots \vee \mathrm{var}_n(\varphi)$. *Concerning (7):* Suppose that $\psi_1, \ldots, \psi_n$ support $\mathrm{var}_1(\varphi) \vee \ldots \vee \mathrm{var}_n(\varphi)$, hence by theorem 3.1(3), (a) $\psi_1, \ldots, \psi_n, \mathrm{var}_1(\neg\varphi) \wedge \ldots \wedge \mathrm{var}_n(\neg\varphi) \vdash_m \bot$ (note that

$\neg \mathrm{var}(\varphi) = \mathrm{var}(\neg\varphi)$). To show that $\psi_1, \ldots, \psi_n$ support φ we show (by theorem 3.1.(3)) that (b) $\psi_1, \ldots, \psi_n, \neg\varphi \vdash_m \bot$. We prove the inference (c) $\psi_1, \ldots, \psi_n, \neg\varphi \vdash_m \mathrm{var}_1(\neg\varphi) \wedge \ldots \wedge \mathrm{var}_n(\neg\varphi)$ by n (AV)-steps and (LC); and (a) + (c) give us (b). The level α/n follows from the general fact that $t(\varphi_1)+\ldots+t(\varphi_n) \geq t(\varphi_1 \vee \ldots \vee \varphi_n)$; and since $t(\mathrm{var}_i(\varphi))$ is equal $t(\varphi)$ for all variants var_i, it follows that $t(\varphi) \geq (t(\mathrm{var}_1(\varphi) \vee \ldots \vee \mathrm{var}_n(\varphi)))/n \geq \alpha/n$. qed

Note that theorem 3.2.(4) does no longer hold if m-entailment is replaced by support. For example, $xRy \vee yRx$ supports xRy, and xRy m-entails and thus supports $xRy \wedge yRx$ (by AV and LC), but $xRy \vee yRx$ does not support $xRy \wedge yRx$. When xRy is interpreted as $x > y$, the degree of truth of $xRy \vee yRx$ approaches 1 but the degree of truth of $xRy \wedge yRx$ equals 0. This examples tells us also that the support-relation is not transitive, though theorem 3.2.(7) states a significant exception.

As a final example, consider the premise $xRy \leftrightarrow \neg yRx$ – that x has relation R to y iff y does not have it to x. In spite of the fact that its universalization, $\forall x \forall y(xRy \leftrightarrow yRx)$, is logically inconsistent, the open formula is m-consistent: since it has the degree of truth 1 when xRy is interpreted as x>y in the domain of natural numbers. An interesting thing about this is that $xRy \leftrightarrow \neg yRx$ m-entails both $xRy \vee yRx$ and $\neg xRy \vee \neg yRx$, and it followings from the forgoing that $xRy \vee yRx$ supports xRy, and $\neg xRy \vee \neg yRx$ supports $\neg xRy$ – hence $xRy \leftrightarrow \neg yRx$ supports both xRy and $\neg xRy$.

4 Measure-entailment and support among conditional a.g.s

4.1 Truthfunctional compounds of conditional a.g.s

In this section we consider m-entailment and support between conditionals of the general form $\varphi \Rightarrow \psi$ and truthfunctional compounds of them, where $\varphi, \psi, \ldots$ range over formulas of the 1st order language excluding $\Rightarrow$. The conditional operator $\Rightarrow$ binds the variables of φ and ψ and, thus, can also be regarded as a dyadic most-quantifier $(M)(\psi/\varphi)$. Hence, the most-quantifier in the language with conditionals is not implicit but *explicit*. Therefore, all standard structural laws for classical consequence are satisfied by the m-entailment relation, provided it is suitably extended to truthfunctional compounds of c.a.g.s – which is the task of this section.

A conditional a.g. with an antecedent which is not logically true is called a *proper* conditional a.g. An absolute a.g., φ, is defined in our conditional language

(the 1st order language plus $\Rightarrow$) as the (improper) conditional a.g. $\top \Rightarrow \varphi$ with tautological antecedent. If '$\top \Rightarrow \varphi$' stands alone, one may write just 'φ' instead of '$\top \Rightarrow \varphi$', but care must be taken if '$\top \Rightarrow \varphi$' occurs in the scope of a truth-functional operator: for example, '$\neg\varphi$' of the absolute language corresponds to '$\top \Rightarrow \neg\varphi$' in the conditional language, which is very different from '$\neg(\top \Rightarrow \varphi)$' (one cannot express the latter assertion without an explicit most-quantifier).

We define the conditional degree-of-truth $t(\varphi \Rightarrow \psi)$ in the standard way as $t(\varphi \wedge \psi)/t(\varphi)$ if $t(\varphi) > 0$, and set to 1 if $t(\varphi) = 0$.[7] The degree-of-falsity $f(\varphi \Rightarrow \psi)$ is defined as $1 - t(\varphi \Rightarrow \psi)$. In the following, $C, C_1, \ldots$ range over conditional a.g.s and $\mathbf{C}, \mathbf{C_1}, \ldots$ over set of them. Moreover, $TC_i, \ldots$ range over arbitrary truth-functional compounds of conditionals and $\mathbf{TC_i}$ over stets of them. M-entailment and support between a finite set of conditional a.g.s and a conditional a.g. is defined as for absolute a.g.s. An extension of the logic of m-entailment for propositional formulas or 'x-formulas' (see sec. 4.2) which allows the conclusion to be a disjunction of conditionals was developed in Adams (1986) and goes as follows: $\mathbf{C} \Vdash_m C_1 \vee \ldots C_n$ iff for all ϵ there exists δ such that for all degree-of-truth functions t, if $t(C') \geq 1 - \delta$ for all $C' \in \mathbf{C}$, then $t(C_i) \geq 1 - \epsilon$ for at least one $C_i (1 \leq i \leq n)$. We call these m-entailments disjunctive m-entailments. Schurz (1997) reformulates them as m-entailments with negated conditionals in the premise set: $\mathbf{C}, \neg C_1 \Vdash_m C_2$ iff for all ϵ there exists δ such that for all degree-of-truth functions t, if $t(C') \geq 1 - \delta$ for all $C' \in C$ and not $t(C_1) \geq 1 - \epsilon$, then $t(C_2) \geq 1 - \epsilon$. This section generalizes these extensions to arbitrary truthfunctional compounds of conditional a.g.s (of the full 1st order language) with help of a recursive definition of 'ϵ-satisfaction' and 'δ-satisfaction'.

Definition 4.1.1. *Let t be a degree-of-truth function.*

(1) t ϵ-satisfies [δ-satisfies] $\varphi \Rightarrow \psi$ iff $t(\varphi \Rightarrow \psi) \geq 1 - \epsilon$ [$\geq 1 - \delta$].
t ϵ-satisfies [δ-satisfies] $\neg TC$ iff t does not δ-satisfy [ϵ-satisfy] TC.
t ϵ-satisfies [δ-satisfies] $TC_1 \vee TC_2$ iff t ϵ-satisfies [δ-satisfies] TC_1 or t ϵ-satisfies [δ-satisfies] TC_2; and likewise for $\wedge$ and $\rightarrow$.

(2) $\mathbf{TC} \Vdash_m TC'$ iff for all ϵ there exists δ such that for all degree-of-truth-functions, if t δ-satisfies all $TC \in \mathbf{TC}$, then t ϵ-satisfies TC'.

Observe the ϵ-δ-switch in the clause for negated conditionals, which corresponds to their 'negativization' by a premise-conclusion-switch. It is straightforward to prove from this definition that m-entailment satisfies all rules for classical

7 A not yet elaborated alternative would consist in the use of primitive conditional degrees-of-truths, similar to Popper functions for (epistemic) probabilities (cf. Hawthorne 1998).

consequence. In theorem 4.1.1.(3) we assume the empty premise set to be interchangeable with 'conditional Verum' $\top \Rightarrow \top$, and the empty conclusion (with zero disjuncts) to be interchangeable with 'conditional Falsum' $\top \Rightarrow \bot$.

Theorem 4.1.1. *Let t be a degree-of-truth function. Then:*

(1) If TC and TC′ are truthfunctionally equivalent, then t ϵ-satisfies [δ-satisfies] TC iff t ϵ-satisfies [δ-satisfies] TC.

(2) (a) $TC_1 \vee TC_2 \Vdash_m TC_3$ *iff* $TC_1 \Vdash_m TC_3$ *and* $TC_2 \Vdash_m TC_3$.
(b) $TC_1 \Vdash_m TC_2 \wedge TC_3$ *iff* $TC_1 \Vdash_m TC_2$ *and* $TC_2 \Vdash_m TC_3$.
(c) $TC_1 \wedge TC_2 \Vdash_m TC_3$ *iff* $TC_1 \Vdash_m \neg TC_2 \vee TC_3$.
(d) $TC_1 \wedge \neg TC_2 \Vdash_m TC_3$ *iff* $TC_1 \Vdash_m TC_2 \vee TC_3$.

(3) For every compound m-entailment $\mathbf{TC} \Vdash_m TC'$ *there exists a finite set of valid disjunctive m-entailments* $\mathbf{C_i} \Vdash_m D_i$ *(with* $1 \leq i \leq n$*, and* D_i *a disjunction of conditional a.g.s) such that* $\mathbf{TC} \Vdash_m TC'$ *is valid iff every m-entailment* $\mathbf{C_i} \Vdash_m D_i$ *is valid.*

Proof. (*Sketch*) To prove (1) one shows, first, that ϵ- and δ-satisfaction is preserved by the equivalence rules of deMorgan, double negation, elimination of $\rightarrow$ by $\vee$, and $\vee\wedge$-distribution. For double negation, e.g., we argue that t ϵ-satisfies $\neg\neg TC$ iff t does not δ-satisfy $\neg TC$ iff t does not not ϵ-satisfy TC iff t ϵ-satisfies TC. Since truthfunctionally equivalent formulas can be transformed into each other by using these the equivalence laws, (1) follows by straightforward induction. The proofs of (2) are straightforward. *Concerning (3):* by conjoining the premises in **TC** and by (1), we first transform the given compound inference into a covalid inference with premise in disjunctive normal form and conclusion in conjunctive normal form. By 2(a,b), we transform this inference into a set of jointly covalid inferences with a conjunction of unnegated or negated conditionals in the premise, and a disjunction of those in the conclusion. We finally eliminate negations by (2c,d) and replace the premise-conjunction by a set, and obtain a jointly covalid set of disjunctive inferences. qed

The result of this section makes it possible to state the following completeness theorems just for disjunctive inferences; its validity for compound inferences follows automatically from theorem 4.1.1. By use of this theorem, also non-infinitesimal degree-of-truth-conditions for compound inferences can be translated into conditions for disjunctive inferences. Note finally that with help of (negated) conditionals we may express the relation of support between absolute a.g.s as follows:

Fact 4.1.1. $\varphi_1, \ldots, \varphi_n$ *supports* ψ *iff* $\top \Rightarrow \varphi_1, \ldots, \top \Rightarrow \varphi_n \Vdash_m \neg(\top \Rightarrow \neg\psi)$.

4.2 The 'essentially propositional' one-variable case

Most work on conditional probability logic has been done for propositional languages. Formulas of a first order language with monadic predicates and exactly one free variable x are called x-formulas in Adams (1974, §4). Their degree-of-truth functions obey not more laws than those for propositional probabilities; the additional laws of truth-functions concern only multi-variable formulas. Therefore, all theorems and proofs for propositional probabilistic logic can be transferred to the logic of conditional a.g.s in the language of x-formulas. The propositional logic of probabilistic conditionals (Adams 1965, 1975, 1977, 1986, 1988) was transferred to x-formulas in Adams (1974, §4) and Schurz (1997). In this section, we restrict our attention to m-entailment between (truth-functional combinations of) conditionals $\varphi \Rightarrow \psi$ where φ, ψ are formulas of the x-formula-language (such as Fx, Gx, Rax, $\exists yRyx$, etc.). A complete calculus for disjunctive inferences consists of the following rules:

Calculus C0: $\mathbf{C} \vdash_{C0} C_1 \vee \ldots \vee C_n$ iff $C_1 \vee \ldots \vee C_n$ is derivable from $\mathbf{C}$ by applying structural rules of classical propositional inference and, in addition:

- (CC) $\varphi \Rightarrow \psi, \psi \Rightarrow \chi \vdash_{C0} \varphi \wedge \psi \Rightarrow \chi$ (Cautious Cut)
- (CM) $\varphi \Rightarrow \psi, \varphi \Rightarrow \chi \vdash_{C0} \varphi \wedge \psi \Rightarrow \chi$ (Cautious Monotony)
- (Or) $\varphi \Rightarrow \chi, \psi \Rightarrow \chi \vdash_{C0} \varphi \vee \psi \Rightarrow \chi$ (Or)
- (RM) $\varphi \Rightarrow \chi \vdash_{C0} (\varphi \wedge \psi \Rightarrow \chi) \vee (\varphi \Rightarrow \neg\psi)$ (Rational monotony)
- (SC) If $\varphi \vdash \psi$, then $\vdash_{C0} \varphi \Rightarrow \psi$ (Supraclassicality)

Calculus **C0** 'minus' rule (RM) coincides with the calculus **P** for *preferential entailment* (cf. Makinson 1994).[8]

Well-known derived rules are of **P** are (Left Logical Equivalence, LLE): If $\vdash \varphi \leftrightarrow \chi$, then $\varphi \Rightarrow \psi \vdash_{C0} \chi \Rightarrow \psi$, (Right Weakening, RW): If $\psi \vdash \chi$, then $\varphi \Rightarrow \psi \vdash_{C0} \varphi \Rightarrow \chi$, (And): $\varphi \Rightarrow \psi, \varphi \Rightarrow \chi \vdash_{C0} \varphi \Rightarrow \psi \wedge \chi$ and (Conditional Proof, CP): $\varphi \wedge \psi \Rightarrow \chi \vdash_{C0} \varphi \Rightarrow (\psi \rightarrow \chi)$.

Theorem 4.2.1. *The following conditions are equivalent ('Π' for 'product'):*

(1) $\mathbf{C} \vdash_m C_1 \vee \ldots \vee C_n$

(2) $\mathbf{C} \Vdash_{C0} C_1 \vee \ldots \vee C_n$

(3) For all degree-of-truth-functions, $\Pi\{f(C_i) : 1 \leq i \leq n\} \leq \Sigma\{f(C') : C' \in \mathbf{C}\}$

8 While calculi **P** and **M** have 'Cut' as their only structural rule, the admission of all structural rules of propositonal logic is important for the completeness of **C0**. Rule (RM) corresponds to 'rational monotony' in the sense of Lehmann and Magidor (1992).

For the proof cf. Adams (1986) and Schurz (1998). A fourth equivalent semantical condition is expressed in terms of *ranked models* for x-formulas (Adams 1986, Schurz 1997), and a fifth equivalent condition is the yielding relation as described in Adams (1986, p. 264), and Schurz (1998, p. 85–87), which in the quantifier-free case gives a semi-tractable decision procedure. Condition (3) establishes non-infinitesimal probability semantics for the logic of c.a.g.s which is the basis of the probabilistically reliable non-monotonic theorem prover developed in Schurz (1997).

As in the case for absolute x-formulas, it holds for conditional x-formulas that if $C_1, \ldots, C_n$ do not m-entail C, then there exists a degree-of-truth function such that $t(C_i)$ is arbitrary close to 1 while $t(C)$ is zero (cf. Adams 1986, p. 261). So m-entailment and support coincide for conditional x-formulas.

4.3 The general 1st-order case

Due to the law of independence, m-entailment between conditional a.g.s with an unrestricted number of variables satisfies the two additional rules (IndE) for 'Independent Expansion' and 'IndC' for 'Independent Contraction' (the rules AV and DC are the same as for absolute a.g.s):

Calculus C1 ($\vdash_{C1}$): The rules of **C0** plus
- (AV) $\varphi \Rightarrow \psi \ \vdash_{C1} \mathrm{var}(\varphi \Rightarrow \psi)$ for every $\mathrm{var}(\varphi \Rightarrow \psi)$
- (DC) $\top \Rightarrow \varphi \vee \psi \ \vdash_{C1} \top \Rightarrow \varphi \vee c_\varphi(\psi)$ for every $c_\varphi(\psi)$
- (IndE) $\varphi \Rightarrow \psi \ \vdash_{C1} \varphi \wedge \chi \Rightarrow \psi$ provided *
- (IndC) $\varphi \wedge \chi \Rightarrow \psi \vdash_{C1} (\varphi \Rightarrow \psi) \vee (\chi \Rightarrow \bot)$ provided *

$*$: χ and (φ, ψ) have no free variables in common.

The right disjunct in the conclusion of (IndC) takes care of the case where $t(\chi) = 0$. Rule (IndE) preserves the degree of truth exactly, and (IndC) preserves the degree of truth from the premise to the first conclusion-disjunct exactly, except when $t(\chi) = 0$ and thus $t(\chi \Rightarrow \bot) = 1$. Unfortunately, it is not possible to transfer the completeness results about absolute a.g.s to conditional a.g.s with more than one variable. One reason for this is the *breakdown* of the law of *conditionalization*: for example, $t(Fx \wedge Gy) = t(Fx) \cdot t(Gy)$, but $t(Fx \wedge Gy/Rxy)$ may differ from $t(Fx/Rxy) \cdot t(Gy/Rxy)$ whenever Rxy's extension is not decomposable into a Cartesian product. Therefore, the unconditional rule (DC) has no conditional counterpart. So far, it is an unsolved problem whether calculus **C1** is complete, and more generally, whether m-entailment between general 1st order c.a.g.s is

recursively axiomatizable at all – even in the case where identity is excluded. Calculus **C1** is nevertheless highly important for practical applications because its rules are correct (further independent correct rules have, so far, not been found). Moreover, for conditionals derived without (DC) there exists a certain standard-derivation, which implies a general lower bound for the degree-of-truth of the conclusion under the additional proviso that (IndC) was not used in the derivation:

Theorem 4.3.1.

(1) If $\mathbf{C} \vdash_{C1} C_1 \vee \ldots \vee C_n$*, then* $\mathbf{C} \Vdash_m C_1 \vee \ldots \vee C_n$.

(2) If $C_1 \vee \ldots \vee C_n$ *is derivable from* **C** *in calculus* **C1** *without (DC), then there exists a derivation in which all (AV) and (IndE) steps are advanced.*

(3) If $C_1 \vee \ldots \vee C_n$ *has a derivation from* **C** *in calculus* **C1** *without (DC) and (IndC), then for all degree-of-truth-functions it holds that* $\Pi\{f(C_i) : 1 \leq i \leq n\} \leq \Sigma\{f(C') : C' \in AVIndE(\mathbf{C})\}$*, where AVInd(***C***) is the set of all conditionals which have been derived from* **C** *by the rules (AV) and (IndE).*

Proof. Theorem 4.3.1.(1): Correctness of (IndE) and (IndC) is obvious; correctness of **C1** follows straightforwardly.

Theorem 4.3.1.(2): It is sufficient to show that the set of conditionals derivable from $AVIndE(\mathbf{C})$ by the rules of $\mathbf{C0} + (IndC)$ is closed under the rules (AV) and (IndE):

For (CC): Assume $\varphi \Rightarrow \psi$ is derived from $\varphi \Rightarrow \chi$ and $\varphi \wedge \chi \Rightarrow \psi$ by (CC). *Concerning (AV):* By induction hypothesis (IH), $\mathrm{var}(\varphi) \Rightarrow \mathrm{var}(\chi)$ and $\mathrm{var}(\varphi) \wedge \mathrm{var}(\chi) \Rightarrow \mathrm{var}(\psi)$ is derivable from AVIndE(**C**) (for arbitrary var), which implies $\mathrm{var}(\varphi) \Rightarrow \mathrm{var}(\psi)$ by one additional step of (CC). *Concerning (IndE):* Assume π shares no free variables with (φ, ψ). By IH, we can derive $\varphi \wedge \pi \Rightarrow \chi^*$ and $\varphi \wedge \chi^* \wedge \pi \Rightarrow \psi$ from AVIndE(**C**), where χ^* is a χ-variant which shares no free variables with π. One additional step of (CC) gives us $\varphi \wedge \chi^* \Rightarrow \psi$, and $\varphi \wedge \chi \Rightarrow \psi$ is an alphabetic variant thereof and hence derivable from it by the forgoing argument. Without any complications the same is demonstrable for the rules (Or), (CM), and (SC). *For (RM):* Assume $(\varphi_1 \wedge \varphi_2 \Rightarrow \psi) \vee (\varphi_1 \Rightarrow \neg\varphi_2)$ is derived from $\varphi_1 \Rightarrow \psi$. *Concerning (AV):* the argument is like that for (CC) above. The case of (IndE) is more involved: we must show that $(\varphi_1 \wedge \varphi_2 \wedge \chi_1 \Rightarrow \psi) \vee (\varphi_1 \wedge \chi_2 \Rightarrow \neg\varphi_2)$ is derivable from AVIndE(**C**), for arbitrary χ_i which have no variables common with $(\varphi_1, \varphi_2, \psi)$. By IH, $\varphi_1 \wedge \chi_1 \wedge \chi_2 \Rightarrow \psi$ is derivable from AVIndE(**C**). By (RM), $(\varphi_1 \wedge \varphi_2 \wedge \chi_1 \wedge \chi_2 \Rightarrow \psi) \vee (\varphi_1 \wedge \chi_1 \wedge \chi_2 \Rightarrow \neg\varphi_2)$ follows. This gives us, by (IndC) and propositional rules, $(\varphi_1 \wedge \varphi_2 \wedge \chi_1 \Rightarrow \psi) \vee (\varphi_1 \wedge \chi_2 \Rightarrow \neg\varphi_2) \vee (\chi_1 \Rightarrow \bot) \vee (\chi_2 \Rightarrow \bot)$. But from $\chi_1 \Rightarrow \bot$ we can derive $\varphi_1 \wedge \varphi_2 \wedge \chi_1 \Rightarrow \psi$, and from $\chi_2 \Rightarrow \bot$ we can derive $\varphi_1 \wedge \chi_2 \Rightarrow \neg\varphi_2$; and

so we obtain (by $\vee$-contraction) $(\varphi_1 \wedge \varphi_2 \wedge \chi_1 \Rightarrow \psi) \vee (\varphi_1 \wedge \chi_2 \Rightarrow \neg\varphi_2)$. *For (IndC)*, we argue in the same way.[9]

Theorem 4.3.1.(3): If (IndC) is not used in the proof, too, then $C_1 \vee \ldots \vee C_n$ is derivable from $AVIndE(\mathbf{C})$ solely by the rules of **C0**, and we can apply theorem 4.2.1 to obtain this result. qed

Theorem 4.3.1 is of importance for reliable theorem-proving in calculus **C1**. It is also of significance for a claim of Weydert (1997, p. 596), which says that a calculus essentially consisting of **C0**+(AV)+(IndE)+(IndC), but without the rule (DC), is complete for 1st order m-entailment. If this were true, then rule (DC) would be derivable from **C0**+(AV)+(IndE)+(IndC). Theorem 4.3.1.(2) implies that if $\top \Rightarrow Rxy \vee Qxy$ were derivable from $\top \Rightarrow Rxy \vee Qxz$ in **C0**+(AV)+(IndE)+(IndC), then it must be derivable from $AVIndE(\top \Rightarrow Rxy \vee Qxz)$ in **C0** + (IndC). It is easily to show that in this derivation all (IndC)-steps are eliminable, because they can only refer to antecedents which have been introduced by (IndE), since all elements of $AV(\top \Rightarrow Rxy \vee Qxz)$ have $\top$ as their antecedent. So if Weydert's claim were true, then $\top \Rightarrow Rxy \vee Qxy$ were derivable from alphabetic variants of $\top \Rightarrow Rxy \vee Qxz$ in **C0**. But this cannot be because $\top \Rightarrow Rxy \vee Qxy$ does not yield $\top \Rightarrow Rxy \vee Qxz$ in the sense of Adams (1975, p. 61). So the rule (DC) is not derivable from **C0**+(AV)+(IndE)+(IndC), hence Weydert's system of rules cannot be complete for m-entailment.

The relation of support among conditional a.g.s is defined as for absolute a.g.s. There exists no equivalent formulation of conditional support in terms of m-inconsistency, i.e., in terms of m-entailment of $\top \Rightarrow \bot$. However, an equivalent formulation exists in terms of ϵ-inconsistency in the sense of (Adams, 1975, p. 51), i.e., in terms of m-entailment of $\varphi \Rightarrow \bot$ from the premises and the denial $\varphi \Rightarrow \neg\psi$ of the supported formula $\varphi \Rightarrow \psi$ (theorem 4.3.2.(2)). Since we have no completeness proof for 1st order m-entailment, we can only give a correctness condition analogous to that of theorem 4.3.1.

Theorem 4.3.2.

(1) $C_1, \ldots, C_n$ *support* $\varphi \Rightarrow \psi$ *iff* $C_1, \ldots, C_n, \varphi \Rightarrow \neg\psi \Vdash_m \varphi \Rightarrow \bot$.

(2) If $C_1, \ldots, C_n, \varphi \Rightarrow \neg\psi \vdash_{C1} \varphi \Rightarrow \bot$, *then* $C_1, \ldots, C_n$ *support* $\varphi \Rightarrow \psi$.

(3) If $\varphi \Rightarrow \bot$ *is has a derivation from* $C_1, \ldots, C_n, \varphi \Rightarrow \neg\psi$ *in* ***C1*** *without (DC) and (IndC), then (i)* $C_1, \ldots, C_n$ *support* $\varphi \Rightarrow \psi$ *at level* $(1/r)$, *where* r *is the number of variants of* $\varphi \Rightarrow \neg\psi$ *used in this derivation, and (ii) for all*

9 Observe why the proof breaks down for (DC): if $\top \Rightarrow \varphi \vee \psi'$ is derived from $\top \Rightarrow \varphi \vee \psi$ by (DC), then by IH for (IndE), $\pi \Rightarrow \varphi \vee \psi$ (with variable-disjoint π) is derivable from AVIndE(C), but $\pi \Rightarrow \varphi \vee \psi'$ cannot be concluded because (DC) does not hold for conditional a.g.s.

degree-of-truth-functions t, $t(\varphi \Rightarrow \psi) \geq (1/r) - (\Sigma\{k_i \cdot f(Ci) : 1 \leq i \leq n\})/r$, where k_i is the number of variants of C_i used in this derivation.

Proof. (1), not-right implies-not-left: $C_1, \ldots, C_n, \varphi \Rightarrow \neg\psi \nvdash_m \varphi \Rightarrow \bot$ implies by definition that there exists ϵ such that for all δ there exists a degree-of-truth function t such that (a) $t(C_1), \ldots, t(C_n) \geq 1 - \delta$, (b) $t(\varphi \Rightarrow \neg\psi) \geq 1 - \delta$, and (c) $t(\varphi \Rightarrow \bot) < 1 - \epsilon$. From (c) it follows that $t(\varphi) > 0$, which together with (b) implies that $t(\varphi \Rightarrow \psi) < \delta$. This implies, by definition of support, that there exists no positive level α at which $C_1, \ldots, C_n$ could support $\varphi \Rightarrow \psi$.

(1), not-left-to-not-right: That $C_1, \ldots, C_n$ does not support $\varphi \Rightarrow \psi$ at any positive level α implies that (a) for every (small) α there exists δ such that for every t, $t(C_i) \geq 1 - \delta$ for all $1 \leq i \leq n$, and $t(\varphi \Rightarrow \psi) < \alpha$, which in turn implies $t(\varphi) > 0$. Assume for *reductio* that $C_1, \ldots, C_n, \varphi \Rightarrow \neg\psi \Vdash_m \varphi \Rightarrow \bot$. This means that (b) for every ϵ exists δ such that for all t, if $t(C_i) \geq 1 - \delta$ for all $1 \leq i \leq n$ and $t(\varphi \Rightarrow \neg\psi) \geq 1 - \delta$, then $t(\varphi \Rightarrow \bot) \geq 1 - \epsilon$ and hence $t(\varphi) = 0$. Putting $\epsilon = \alpha$, (b) implies that (c) for every small α there exists δ such that for every t, if $t(C_i) \geq 1 - \delta$ for all $1 \leq i \leq n$ and $t(\varphi) > 0$, then $t(\varphi \Rightarrow \neg\psi) < \alpha$ and hence, because of $t(\varphi) > 0$, $t(\varphi \Rightarrow \psi) \geq \alpha$. But (c) contradicts (a).

(2). follows from theorem 4.3.2.(1) and correctness of **C1**.

Concerning (3): Given the assumption, theorem 4.3.1.(3) implies that (i) for all t, $f(\varphi \Rightarrow \bot) \leq \Sigma\{k_i \cdot C_i : 1 \leq i \leq n\} + r \cdot f(\varphi \Rightarrow \neg\psi)$, with k_i and r as explained. If $t(\varphi) = 0$, then $t(\varphi \Rightarrow \psi) = 1$ and our claim is trivially satisfied. Else $t(\varphi) > 0$, and then $f(\varphi \Rightarrow \bot) = 1$, $f(\varphi \Rightarrow \neg\psi) = t(\psi \Rightarrow \psi)$, which gives our claim by (i).

qed

We conclude with an example: $Fx \wedge Fy \Rightarrow Rxy \vee Ryx$ supports $Fx \wedge Fy \Rightarrow Rxy$ at level 1/2, because two AV-instances of the denied conclusion, namely $Fx \wedge Fy \Rightarrow \neg Rxy$ itself and $Fy \wedge Fx \Rightarrow \neg Ryx$ are necessary to derive $Fx \wedge Fy \Rightarrow \bot$. In the following sections we focus on the support of antecedents of conditional a.g.s.

4.4 Entailment of conditional a.g.s with supported antecedent and restricted 1st order completeness theorems

For present purposes, the most significant fact about a conditional a.g. is its connection with an absolute a.g. by the following relation:

Fact 4.4.1. $\varphi \rightarrow \psi \Vdash_m (\varphi \Rightarrow \psi) \vee (\top \Rightarrow \neg\varphi)$, *or equivalently*

$$\neg(\top \Rightarrow \neg\varphi), \varphi \rightarrow \psi \Vdash_m \varphi \Rightarrow \psi$$

with $f(\varphi \Rightarrow \psi) \leq f(\varphi \rightarrow \psi)/t(\varphi)$ *for all degree-of-truth functions* t.

This follows by rule (RM) from theorem 4.3.1. Note that in this special case, the inequality can be replaced by an equality. Fact 4.4.1 is exploited in the following:

Theorem 4.4.1. *If a set of (absolute or conditional) a.g.s m-entails a material implication $\varphi \rightarrow \psi$ and at the same time supports its antecedent φ, then this set m-entails the corresponding conditional a.g. $\varphi \Rightarrow \psi$.*

Proof. This is a direct consequence of facts 4.1.1 and 4.4.1. qed

A simple application of the foregoing is that while $xRy \vee yRx$ does not m-entail $\neg xRy \Rightarrow yRx$[10], $xRy \vee yRx$ together with $\neg xRy \vee \neg yRx$ does m-entail it. Thus, clearly $xRy \vee yRx$ by itself m-entails the material conditional $\neg xRy \rightarrow yRx$, and therefore so does the combination of $xRy \vee yRx$ and $\neg xRy \vee \neg yRx$. Moreover, since both $\neg xRy$ and $\neg yRx$ are alphabetic variants of $\neg xRy$, the disjunction $\neg xRy \vee \neg yRx$ supports $\neg xRy$ (by theorem 3.2.6). Hence theorem 4.4.1 guarantees that $xRy \vee yRx$ together with $\neg xRy \vee \neg yRx$ m-entails $\neg xRy \Rightarrow yRx$.

This is our first application in which absolute a.g.s are shown to m-entail a conditional a.g. In the area of conditional propositions, $\psi \Rightarrow \varphi$, where ψ and φ are sentences, this can only happen if either φ is a logical consequence of ψ or both ψ and φ are entailed by the premises.[11] Now we see that this can happen with conditional a.g.s $\psi \Rightarrow \varphi$ with more than one free variable, when the corresponding material conditionals, $\psi \rightarrow \varphi$, are m-entailed, and the premises only *support* their antecedents.

Generally speaking, theorem 4.4.1 becomes useful whenever a given set of a.g.s, e.g. a scientific theory T, m-entails a material implication $\varphi \rightarrow \psi$ and we want to infer from this fact something about the corresponding conditional a.g. $\varphi \Rightarrow \psi$. For example, if we want to predict ψ-instances on the basis of φ-instances we must ensure that the conditional degree of truth of ψ given φ is sufficiently high; a high degree of truth of $\varphi \rightarrow \psi$ is not sufficient for this purpose (recall the relative-son example in the introduction). We may utilize theorem 4.4.1 in two ways. *First*, it may be that theory T also supports the antecedent of the m-entailed implication, whence we can conclude that the conditional a.g. is m-entailed. *Second*, it may be that T together with some plausible *default* premises (such that there exist more than just a few objects) supports the antecedent. Then T enriched by these default premises m-entails the conditional a.g. In the following

10 Thus, if xRy is interpreted as $\neg x = y$ in the domain of natural numbers, then $xRy \vee yRx$ is true for all pairs $\langle x, y \rangle$ with $\neg x = y$, but the proportion of pairs $\langle x, y \rangle$ that satisfy yRx out of all those that satisfy $\neg xRy$ is 0.

11 This is essentially the *factuality-conditionality property* stated on p. 175 of Adams (1988).

sections on applications we make use of both possibilities. But before that, we demonstrate how certain restricted completeness theorems can be derived from theorem 4.4.1. We can prove 1st order completeness for all conclusions of the form $(\varphi \Rightarrow \psi) \vee (\top \Rightarrow \neg\varphi)$, and for all conditional conclusions with supported antecedent (see theorem 4.4.2). Important for the proof is lemma 4.4.1, which tells us that an absolute a.g. which is m-entailed by a set of conditional premises $\mathbf{C}$ is already m-entailed by the set of material counterparts $\mathbf{C}^{\rightarrow} =_{df} \{\varphi \rightarrow \psi : \varphi \Rightarrow \psi \in \mathbf{C}\}$ of these premises.

Lemma 4.4.1.
(1) $\mathbf{C} \Vdash_m (\varphi \Rightarrow \psi) \vee (\top \Rightarrow \neg\varphi)$ *iff* $\mathbf{C} \Vdash_m \top \Rightarrow (\varphi \rightarrow \psi)$
(2) $\mathbf{C} \Vdash_m \top \Rightarrow \varphi$ *iff* $\mathbf{C}^{\rightarrow} \Vdash_m \varphi$
(3) $\mathbf{C} \Vdash_m (\varphi \Rightarrow \psi) \vee (\top \Rightarrow \neg\varphi)$ *iff* $\mathbf{C}^{\rightarrow} \Vdash_m \varphi \rightarrow \psi$

Proof. (1), left-to-right, holds because both $\varphi \Rightarrow \psi$ and $\top \Rightarrow \neg\varphi$ m-entail $\top \Rightarrow (\varphi \rightarrow \psi)$. (1), right-to-left, follows from fact 4.4.1.

Concerning (2): The right-to-left direction holds because conditionals m-entail their material counterparts (by the rule CP). We prove the left-to-right direction by contraposition. Assume $\mathbf{C}^{\rightarrow} \nVdash_m \varphi$, where $x_1, \ldots, x_m$ are φ's free variables. By theorem 2.1.3 this implies that there exists a model M such that all implications $\psi \rightarrow \psi'$ in $\mathbf{C}^{\rightarrow}$ are satisfied in M under all valuations which assign distinct values in $D \cup D_1 \cup \ldots \cup D_m$ to distinct free variables, and φ is false in M under all valuations v such that $v(x_i) \in D_i$ for $1 \leq i \leq m$. This implies, by the argument of Adams (1974, p. 8), that if the size of the domain goes to infinite then $t(\psi \rightarrow \psi')$ approaches 1, i.e., $f(\psi \rightarrow \psi')$ approaches 0, while $t(\varphi)$ cannot be greater that $1 - m^{-m}$. Now note that (i) $f(\psi \Rightarrow \psi') = f(\psi \rightarrow \psi')/t(\psi)$. We argue as follows. In the limit model, either (ii) $t(\psi) = 0$, or (iii) $t(\psi) > 0$. In case (ii), $t(\psi \Rightarrow \psi') = 1$ holds by definition. And in case (iii), $f(\psi \Rightarrow \psi') = 0$ and hence $t(\psi \Rightarrow \psi') = 1$ follows from (i) and $f(\psi \rightarrow \psi') = 0$. In both cases we obtain a limit model which gives all conditionals in $\mathbf{C}$ degree-of-truth 1, but $\top \Rightarrow \varphi$ a degree-of-truth $\leq 1 - m^{-m}$. So $\mathbf{C} \nVdash_m \top \Rightarrow \varphi$.[12]

(3) follows from lemma 4.4.1.(1+2).

qed

Theorem 4.4.2 ((Restricted 1st order Completeness for C1):).
(1) $\mathbf{C} \vdash_{C1} (\varphi \Rightarrow \psi) \vee (\top \Rightarrow \neg\varphi)$ *iff* $\mathbf{C} \Vdash_m (\varphi \Rightarrow \psi) \vee (\top \Rightarrow \neg\varphi)$
(2) $\mathbf{C} \vdash_{C1} (\varphi \Rightarrow \psi)$ *iff* $\mathbf{C} \Vdash_m (\varphi \Rightarrow \psi)$, *provided* $\mathbf{C}$ *supports* φ.

12 Observe that this argument breaks down as soon as the conclusion is a proper conditional $(\varphi^* \Rightarrow \varphi)$, because from the fact that $t(\varphi^* \rightarrow \varphi)$ is small we cannot conclude that $t(\varphi^* \Rightarrow \varphi)$ is small.

Proof. Concerning (1): Left-to-right holds by correctness of **C1**. *Right-to-left:* $\mathbf{C} \Vdash_m (\varphi \Rightarrow \psi) \vee (\top \Rightarrow \neg\varphi)$ implies $\mathbf{C}^{\rightarrow} \Vdash_m \varphi \rightarrow \psi$ by lemma 4.4.1.(3), which implies $\mathbf{C}^{\rightarrow} \vdash_m \varphi \rightarrow \psi$ by theorem 2.1. From **C** we obtain $\mathbf{C}^{\rightarrow}$ by applying the rule (CP), and from $\varphi \rightarrow \psi$ we obtain $(\varphi \Rightarrow \psi) \vee (\top \Rightarrow \neg\varphi)$ by the rule (RM). This gives us $\mathbf{C} \vdash_{C1} (\varphi \Rightarrow \psi) \vee (\top \Rightarrow \neg\varphi)$.

Concerning (2): Left-to-right hold by correctness of **C1**. *Right-to-left:* $\mathbf{C} \Vdash_m (\varphi \Rightarrow \psi)$ implies $\mathbf{C} \Vdash_m (\varphi \Rightarrow \psi) \vee (\top \Rightarrow \neg\varphi)$ which implies (i) $\mathbf{C} \vdash_{C1} (\varphi \Rightarrow \psi) \vee (\top \Rightarrow \neg\varphi)$ by theorem 4.4.1.(1). That **C** supports φ implies that $\mathbf{C}, \top \Rightarrow \neg\varphi \Vdash_m \top \Rightarrow \bot$ by theorem 4.3.2.(1), which implies by lemma 4.4.1.(2) that $\mathbf{C}^{\rightarrow}, \neg\varphi \Vdash_m \bot$. This in turn implies that $\mathbf{C}^{\rightarrow}, \neg\varphi \vdash_m \bot$ by theorem 3.1. By applying (CP)-steps to **C** we derive $\mathbf{C}^{\rightarrow}$, which gives us $\mathbf{C}, \top \Rightarrow \varphi \vdash_{C1} \top \Rightarrow \bot$ and hence (ii) $\mathbf{C} \vdash_{C1} \neg(\top \Rightarrow \neg\varphi)$. Now, (i) and (ii) give us $\mathbf{C} \vdash_{C1} \varphi \Rightarrow \psi$.

qed

5 Support in theories of ordering

Theories of *strict ordering*, e.g., of preference between 'options', are commonly based on a binary relation xPy ('x is preferred to y') that satisfies two axioms:[13]

- ASYMM $xPy \rightarrow \neg yPx$, and
- NEGTRANS (negative transitivity) $xPz \rightarrow (xPy \vee yPz)$.[14]

Familiar laws of ordering such as

- TRANS $(xPy \wedge yPz) \rightarrow xPz$

are then deduced from (the universal closure of) these axioms by the principles of standard first-order logic.

Now, although we shall not bother to prove it here, most theorems of theories of strict ordering such as TRANS are not only derivable from but are m-entailed by ASYMM and NEGTRANS.[15] The problem is that the corresponding proper conditional a.g.s are not m-entailed. E.g.,

- ASYMM$\mathbf{C}$ $xPy \Rightarrow \neg yPx$, and
- NEGTRANS$\mathbf{C}$ $xPz \Rightarrow (xPy \vee yPz)$

13 These relations are essentially strict weak orders (or strict quasi-orders) as characterized in theorem 1.3, p. 33, of Roberts (1979). The equivalence classes of a strict quasi-ordering, obtained from the equivalence relation $xEy =_{df} \neg xPy \wedge \neg yPx$, are strictly ordered.

14 This is so called here because it is equivalent to the transitivity of $\neg P$; i.e., to $(\neg xPy \wedge \neg yPz) \rightarrow \neg xPz$.

15 A significant exception is IRREFLEXIVITY, $\neg xPx$, which, because it involves a repeated variable in its atomic constituent, is measure-independent of the other axioms.

are not m-entailed by ASYMM and NEGTRANS. This is obvious in models $M = \langle D, I\rangle$, in which D is large but xPy is interpreted as $x = y$. Then $xPy \rightarrow \neg yPx$ holds for all $x \neq y$ in D, but $t(xPy \Rightarrow \neg yPx) = t(xPy \wedge \neg yPx)/t(xPy) = 0$.[16]

Similar counterexamples show that ASYMM**C** together with NEGTRANS do not m-entail NEGTRANS**C**. The upshot is that not only do the 'absolute' versions of the laws of approximate orderings not m-entail their proper conditional counterparts, but none of these together with any finite subset of the proper conditional counterparts m-entails the conditional counterparts of the remaining ones. We cannot 'finitely axiomatize' the proper conditional theory of approximate ordering relations. But should we want to?

The counterexample to the derivability of ASYMM**C** from ASYMM and the other absolute ordering laws, in which xPy is interpreted as x=y, is an 'ordering in name only', since it wouldn't be recognized as an ordering in any intuitive sense. Intuitive orderings are like those of numbers, real or rational, or like ages of people, and these satisfy further conditions. One that we will focus on is approximate trichotomy (also called weak connectedness):

– TRICHOT $x \neq y \rightarrow (xPy \vee yPx)$.

Another one is the 'default' assumption that the domain contains more than one or just a few objects. We may express this condition by first order sentences which assert that there exists at least k objects,

– MINk $\exists x_1 \ldots \exists x_k \wedge \{x_i \neq x_j : 1 \leq i < j \leq k\}$.

Observe that MINk is a sentence and hence not approximately true but either true or false; yet it will play a role it m-entailment or support-relations. A stronger assumption which m-entails MINk for every k would be $x \neq y$, which m-entails that the domain is potentially infinite. So we could replace MINk by $x \neq y$ in theorem 5.1 below, but we do not need this stronger assumption. If we add these two conditions to TRANS we obtain a fundamental theorem. Let an *n-x-ordering formula* (in the binary predicate P) be a formula of the form $x_{i_1}Px_{i_2} \wedge x_{i_2}Px_{i_3} \ldots \wedge x_{i_{n-1}}Px_{i_n}$, where $x_{i_1}, x_{i_2}, \ldots, x_{i_n}$ is some ordering of $x_1, \ldots, x_n$. Moreover, let $\mathrm{Dist}(x_1, \ldots, x_n)$ be the formula $\bigwedge\{x_i \neq x_j : 1 \leq i < j \leq n\}$ asserting that the x_i denote pairwise distinct objects, and let $\mathrm{DIST}(\varphi) =_{df} \mathrm{DIST}(x_1, \ldots, x_n)$, where $x_1, \ldots, x_n$ are the free variables in φ.

16 Worse, combinations of ASYMM**C**, NEGTRANS**C**, and TRANS**C** do not measure-entail any of the 'higher order transitivity laws', $\mathrm{TRANS}_n\mathbf{C}$ $(x_1Px_2 \wedge \ldots \wedge x_{n-1}Px_n) \Rightarrow x_1Px_n$, and simple counterexamples show that ASYMM**C**, NEGTRANS**C**,$\mathrm{TRANS}_1\mathbf{C}$, …,$\mathrm{TRANS}_n\mathbf{C}$ do not measure-entail $\mathrm{TRANS}_{n+1}\mathbf{C}$.

Lemma 5.1.

(1) TRANS, TRICHOT $\Vdash$ *DIST*$(x_1, \ldots, x_n) \rightarrow \bigvee\{\varphi : \varphi$ *an n-x-ordering formula*$\}$

(2) Provided $k \geq n$, *MINk supports DIST*$(x_1, \ldots, x_n)$ *at level* $(k \cdot (k-1) \cdot \ldots \cdot (k-n+1))/k^n$.

(3) Provided $k \geq n$, $\{$*TRANS, TRICHOT, MINk*$\}$ *supports* $\bigvee\{\varphi : \varphi$ *an n-x-ordering formula*$\}$.

Proof. (1): TRICHOT and DIST$(x_1, \ldots, x_n)$ imply for each pair of variables x_i, x_j, that $(x_i P x_j \vee x_j P x_i)$ will hold. If we apply distribution to the conjunction of these disjunctions and eliminate those contradicting TRANS, we obtain the disjunction of all possible n-x-orderings which is implied by TRANS, TRICHOT and DIST$(x_1, \ldots, x_n)$.[17]

(2): If t(MINk) = 1, then the domain has at least k objects. Then given $n \leq k$, the number of n-tuples of distinct objects among k objects is $k \cdot (k-1) \cdot \ldots \cdot (k-n+1)$, hence its proportion among all k^n n-tuples is $(k \cdot (k-1) \cdot \ldots \cdot (k-n+1))/k^n$. Since this function increases with increasing k, it gives us a lower bound of the number of n-tuples of distinct objects among at least k objects, and thus, a level of support.

(3): $\{$TRANS, TRICHOT, MINk$\}$ m-entail DIST$(x_1, \ldots, x_n) \rightarrow \bigvee\{\varphi : \varphi$ an n-x-ordering formula$\}$ (by (1)) and support DIST$(x_1, \ldots, x_n)$ (by (2)); so they support $\bigvee\{\varphi : \varphi$ an n-x-ordering formula$\}$ by theorem 3.2.(5).

qed

Theorem 5.1.

(1) Provided $k \geq n$, $\{$*TRANS, TRICHOT, MINk*$\}$ *supports every n-x-ordering formula at level* $(k \cdot (k-1) \cdot \cdots \cdot (k-n+1))/k^n . n!$

(2) Assume φ *is a truthfunctional combination of open atomic P-formulas and has free variables* $x_1, \ldots, x_n$. *Then, if* φ *is consistent with* $\{$*ASYMM, TRANS, TRICHOT, DIST*$(x_1, \ldots, x_n)\}$, *then* φ *is supported by* $\{$*ASYMM, TRANS, TRICHOT, MINk*$\}$, *provided* $k \geq n$.

(3) If $\varphi \rightarrow \psi$ *is m-entailed by* $\{$*TRANS, TRICHOT, ASYMM*$\}$ *and* φ *satisfies the conditions of theorem 5.1.(2), then* $\varphi \Rightarrow \psi$ *is m-entailed by* $\{$*TRANS, TRICHOT, ASYMM, MINk*$\}$.

(4) If $\varphi \rightarrow \psi$ *is m-entailed by* $\{$*TRANS, TRICHOT, ASYMM*$\}$ *where* φ *is an n-x-ordering formula, then* $\varphi \Rightarrow \psi$ *is m-entailed by* $\{$*TRANS, TRICHOT, ASYMM*$\}$.

17 Note that without DIST$(x_1, \ldots, x_n)$ the implication is invalid: if all variables denote the same individual, than no n-x-ordering is satisfied by this valuation.

Proof. *(1)* is a direct consequence of lemma 5.1.(3), theorem 3.2.(7) and the fact that there exist $n!$ n-x-ordering formulas, whose disjunction is supported at the level specified in lemma 5.1.(2).

Concerning (2): Given the assumptions about φ, then φ holds in at least one model-valuation pair (M, v) satisfying TRANS, TRICHOT, ASYMM, where distinct values are assigned to φ's variables. In the submodel M_n of M restricted to the n objects $v(x_1), \ldots, v(x_n)$, exactly one n-x-ordering formula ψ will be satisfied, and ψ determines the interpretation of P in M_n modulo isomorphism; so ψ must imply φ in the presence of TRANS, TRICHOT and ASYMM. (Or to argue syntactically: TRANS, TRICHOT, ASYMM and ψ logically imply either x_iPx_j or its negation, for $1 \leq i < j \leq n$; hence they must imply φ, because they don't imply φ's negation.) Therefore, TRANS, TRICHOT, ASYMM, MINk logically imply and thus m-entail $\psi \rightarrow \varphi$, and at the same time they support ψ by theorem 5.1.(1), so they support φ by theorem 3.2.(5).

(3) follows directly from theorems 5.1.(2) and 4.4.1.

Concerning (4): We argue that for all ϵ there exists δ such that for all degree-of-truth functions t with underlying model M, if t(TRANS), t(TRICHOT), t(ASYMM) $\geq 1 - \delta$, then $t(\varphi \Rightarrow \psi) \geq 1 - \epsilon$. Let ϵ be given. *Argument 1:* There exists δ such that for all models M which have at least n objects, our claim holds. This follows from theorem 5.1.(3), since all these models satisfy MINn. *Argument 2*: For the same δ as in argument 1, our claim is also satisfied for all models M which have less than n objects, because by the presence of ASYMM, the x-n-ordering formula φ will be false in these models for all valuations of its variables. Therefore $t(\varphi)$ will be zero and hence $t(\varphi \Rightarrow \psi)$ will be 1 in all of these models. qed

In theorem 5.1.(2), the consistency with ASYMM and DIST$(x_1, \ldots, x_n)$ is important. For example, xPx is not consistent with ASYMM and hence not entailed by any x-n-ordering; $\neg xPx$ is consistent with ASYMM and entailed by every x-n-ordering; $xPy \vee yPy$ is equivalent with the n-x-ordering xPy; finally, $\neg xPy \wedge \neg yPx$ is not consistent with DIST(x, y), it is equivalent with $x = y$ and thus not entailed by any n-x-ordering.

Applying our result to approximate orderings, we use the fact that ASYMM, TRICHOT and TRANS logically entail all of the standard laws of ordering of the form $\varphi \rightarrow \psi$, where φ and ψ are in the language of P. In most cases, the antecedents of these laws satisfy the conditions of theorem 5.1.(3), and therefore the axioms together with the 'default premise' MINk also m-entail the conditional versions of these laws, $\varphi \Rightarrow \psi$.

The foregoing is illustrated by one of the laws of R. D. Luce's theory of *semiorders* (Luce, 1956):

– SEMIORD $(xPy \wedge zPw) \rightarrow (xPw \vee zPy)$.[18]

SEMIORD is *m*-entailed by TRICHOT and TRANS, and given ASYMM, TRICHOT, TRANS and DIST(x, y, z), its antecedent $xPy \wedge zPw$ is equivalent to the disjunction of the six 4-x-orderings $xPyPzPw$, $xPzPyPw$, $xPzPwPy$, $zPxPyPw$, $zPxPwPy$, and $zPwPxPy$, all of which imply $xPy \wedge zPw$. Hence according to Theorem 5.1.(3), ASYMM, TRICHOT, TRANS and Mink with $k \geq 3$ support $xPy \wedge zPw$, and therefore they m-entail:

– SEMIORD**C** $(xPy \wedge zPw) \Rightarrow (xPw \vee zPy)$.

In some cases, the antecedent of a law of ordering is itself an n-x-ordering formula. This is the case, e.g., for ASYMM, NEGTRANS, TRANS, and all higher-order transitivity laws TRANS$_n$ (cf. fn. 16). Theorem 5.1.(4) tells us that the conditional versions of all of these laws are m-entailed by ASYMM, TRANS and TRICHOT even without a default premise about the size of the domain. Also, note that the m-consistent law $xPy \leftrightarrow \neg yPx$ considered previously is also m-entailed by TRICHOT and TRANS, and the present results apply to it. It is equivalent to the conjunction of $xPy \rightarrow \neg yPx$ and $\neg xPy \rightarrow yPx$, and since the antecedents of both of these are n-x-formulas, their properly conditionalized versions, $xPy \Rightarrow \neg yPx$ and $\neg xPy \Rightarrow yPx$, are also measure-entailed by ASYMM, TRICHOT and TRANS.

However, contrary to the impression that may be gained from the foregoing, it should be noted that not all laws of ordering have antecedents that are consistent with ASYMM, TRICHOT, TRANS and DIST$(x_1, \ldots, x_n)$. An important example is

– SUBSTINDIFF $(\neg xPy \wedge \neg yPx \wedge xPz) \rightarrow yPz$,

which is a law of simple orderings. But while this law is m-entailed by TRICHOT and TRANS, its antecedent is not consistent with TRICHOT, TRANS and DIST(x, y). So $\neg xPy \wedge \neg yPx \wedge xPz$ is not implied by any 3-ordering, and its proper conditionalization,

– SUBSTINDIFF**C** $(\neg xPy \wedge \neg yPx \wedge xPz) \Rightarrow yPz$

18 This axiom together with $(xPy \wedge yPz) \rightarrow (xPw \vee wPz)$ replaces NEGTRANS in Luce's theory, and all of this theory's laws follow from ASYMM and NEGTRANS. Hence the results of the present section apply to all of these laws. However, TRICHOT is not a law of this theory, its axioms do not support it, and therefore Luce's theory does not measure-entail laws like SEMIORD**C**.

is not *m*-entailed by them. This shows that the theory of *approximate equivalence*, of which the theory of similarity is one application, must be studied independently.[19]

In the next section we consider another empirical theory which is less widely known than theories of ordering, but has attained some prominence in axiomatic anthropology. This concerns balanced structures (Harary and Per Hage, 1983, p. 44 ff).

6 Deductive support in theory of balanced structures

These structures are exemplified by friendship relations satisfying laws F1 – F4 below. The laws are stated both informally and formally; symbolizing "x is a friend of y" as "xFy", and assuming that an enemy is anyone who is not a friend, they are:

- F1 Friends of friends are friends: $(xFy \wedge yFz) \rightarrow xFz$,
- F2 Friends of enemies are enemies: $(xFy \wedge \neg yFz) \rightarrow \neg xFz$,
- F3 Enemies of friends are enemies: $(\neg xFy \wedge yFz) \rightarrow \neg xFz$,
- F4 Enemies of enemies are friends: $(\neg xFy \wedge \neg yFz) \rightarrow xFz$.

Note that F4 implies xFx, i.e. reflexivity of the friendship relation. Obviously these laws are at best crude approximations even in the most polarized of societies[20], and one wants to see whether their proper conditional counterparts can be expected to hold. In fact, with the addition of an axiom of symmetry,

- SYMM $xFy \rightarrow yFx$,

it can be shown that the proper conditional forms of F1 and SYMM all follow.

---

19 That the proper conditional versions of laws of indifference, or equivalence, are poor approximations in empirical realizations of theories of ordering, such as in application to preferences, is a major motivation for modifying these theories, *e.g.*, as in the theory of semiorders.

20 Interpreting "xFy" as "x repels y", these laws are better approximations in application to electrostatic charges, although F4 is still a bad approximation since it seems to preclude the existence of uncharged or neutral particles. Realism would require weakening the 'extreme polarization' assumption to allow for this, and in the sociological case, to allow for 'multiple cliques'. Allowing for neutrals, enemies are simply another class and not just not friends, and the relation 'x is an enemy of y' must be symbolized independently, *e.g.*, as "xEy". If "$\neg F$" is replaced by "E" throughout, axioms F1 – F4 and SYMM below remain valid, but they no longer entail F4a, namely that two among any three persons must be friends. If neutrals are allowed then all persons could be neutral and neither friends nor enemies. Then the proper conditional versions of these axioms would not follow, since F4a is essential to deriving them.

That the proper conditional form of SYMM follows from the axioms is obvious, since F4 is equivalent to

– F4a $xFy \vee yFz \vee zFx$

all of whose disjuncts are alphabetic variants of xFy. Therefore according to theorem 3.2.(6), F4a supports xFy at level 1/3. And, given that xFy is the antecedent of SYMM, theorem 4.4.1 implies that

– SYMM**C** $xFy \Rightarrow yFx$

is m-entailed by the axioms. The more difficult thing is to show that SYMM together with F1 – F4 support the antecedent of F1, and therefore m-entail Fl's proper conditional form. This can be shown by direct deduction, but this lengthy task would be more suitable for a computer-program in resolution-refutation than for a human brain.[21] We present a proof based on a general theorem and on intuitive graph-theoretical reasoning. Recall that $M = \langle D, I \rangle$ is a model for a formula φ iff M satisfies φ for *all* valuations of φ's free variables.

Theorem 6.1. *Suppose all formulas in $L \cup \{\varphi\}$ (where intuitively L is a set of 'laws' and φ an a.g.) are without quantifiers, identity and constants, and have $x_1, \ldots, x_n$ as their free variables. The models of our language interpret all predicates in $L \cup \{\varphi\}$. Then: if φ is satisfiable in every model of L with exactly $m \geq n$ objects and an injective valuation function (assigning distinct objects to distinct variables), then L supports φ at level $1/m!$.*

Proof. Let $\mathrm{Var}_m(L)$, be the set of all variants of the elements of L in m given variables, and likewise for $\mathrm{Var}_m(\varphi)$, $\mathrm{Var}_m(\neg\varphi)$. Every model of L verifies all elements of $\mathrm{Var}_m(L)$ and, by the assumptions of theorem 6.1, verifies at least one element of $\mathrm{Var}_m(\varphi)$ with help of an injective valuation function. Because L and φ are free of quantifiers, identity and constants, every possible truth-valuation on the atomic subformulas of $\mathrm{Var}_m(L)$ is realized by exactly one model of $\mathrm{Var}_m(L)$ with m objects and an injective valuation function. Therefore, every truth-function verifying each element of $\mathrm{Var}_m(L)$ verifies one element of $\mathrm{Var}_m(\varphi)$. Hence, $\mathrm{Var}_m(L) \cap \mathrm{Var}_m(\neg\varphi)$ is truthfunctionally inconsistent. So, by theorem 3.1 ($1 \Leftrightarrow 3 \Leftrightarrow 6$), L supports φ at level $1/m!$ (since m! is the number of variants in m variables). qed

21 The four axioms F1–F4 and the negated antecedent of F1, $\neg xFy \vee \neg yFz$, have each $3! = 6$ variants in $\{x, y, z\}$, and from these 30 premises (clauses) we must derive a contradiction. Already at the first layer, there exist hundreds of possible resolution steps.

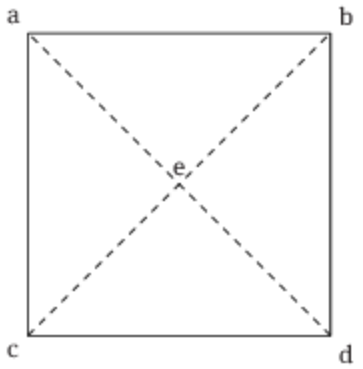


Fig. 1

We apply theorem 6.1 to our problem by identifying φ with $xFy \wedge yFz$ and L with the system of axioms of balance, namely F1–F4 and SYMM. So $n = 3$, and we let $m = 5$, i.e., we consider graphs $\mathcal{G}$ that have 5 'vertices', a, b, c, d, e, as shown figure 1.

Focusing first just on a, b, and c, according to F4a at least two of these must be friends, and therefore be connected by the friendship relation. Assume that a and b are friends, as shown by the heavy horizontal line that connects them in the figure.

Now consider the triple b, c, and d, at least two of which must also be friends. If either b and c or b and d are connected, then $\mathcal{G}$ has a chain of 3 connected vertices, either a,b,c or a,b,d, and therefore $xFy \wedge yFz$ must be satisfied in $\mathcal{G}$ by the assignment either of values a, b, and c or of values a, b, and d to x, y, and z, respectively. The only way that $\mathcal{G}$ might not have such a chain would be for c and d to be friends, as shown in the figure. But then, what about the triple a, c, and e, at least two of whose vertices must also be connected? This simple inspection shows that no matter how connections are graphed between the 5 vertices in $\mathcal{G}$, if at least 2 out three of them are connected then there must be a chain of 3 of them that are connected. Clearly, this is true of any of the finite number of graphs $\mathcal{G}$ on the 5 vertices that is a model of $\mathcal{L}$, and therefore any such graph must have a chain of 3 connected vertices.

Given theorem 6.1, it follows that the formula $\varphi = xFy \wedge yFz$ must be supported by F1–F4, whence by theorem 4.4.1,

– **F1C:** $(xFy \wedge yFz) \Rightarrow xFz$

is m-entailed by SYMM and F1–F4.

That the antecedents of the remaining laws, F2–F4, are not supported by F1–F4, is seen by considering models of F1–F4 where all individuals are friends of each other (and, of course, also friends of themselves). The degree of truth of $\neg xFy$ is zero is these models, and therefore the degree of truth of the antecedents of F2–F4 is zero, too. However, this does not imply that the conditional versions of F2–F4 are not m-entailed by SYMM and F1–F4, because in models where all individuals are friends of each other, the antecedents of F2–F4 have zero degree of truth so that their conditional versions are trivially satisfied. In fact, detailed considerations make is plausible that also F2**C**, F3**C** and F4**C** are *m*-entailed; we leave this as an open conjecture.

Two remarks may be made in concluding this section. First, it shows that in spite of appearances, rules F1–F4 are not symmetrical between friends and enemies. It is possible according to these maxims for a society to be entirely peaceful and friendly, but not for there to be a 'war of all against all'. The key 'desymmetrizing rule' is axiom F4, which entails that no matter how numerous enemies may be, there must at least be some friends, though no similar principle applies to enemies.

Second, that axiom F4a supports xFy at level 1/3 is really a graph-theoretical fact, which shows a connection between the theory of support and pure graph theory. But what is important in this connection is not the mere fact that F4a supports xFy, but rather the degree to which it supports it. Roughly, that F4a supports xFy at level 1/3 means that xFy's degree of truth can be guaranteed to be 'arbitrarily close to at least 1/3' by requiring F4a to have a degree of truth sufficiently close to 1. This requires that the proportion of 'errors' in the formula $xFy \vee yFz \vee zFx$, or equivalently, the proportion of triples in the graph of the relation F that do not satisfy the formula, should be sufficiently close to 0. Thus, we are dealing with graphs with 'small errors' – but not with *random* errors, or *random* graphs. However randomness might be defined in cases like the present one, the assumption that errors are random is as much of an idealization as the assumption that there are no errors. Our purpose here is to avoid idealizing assumptions in reasoning about generalizations, and that applies just as much to randomness as to 'perfect exceptionlessness'.

7 Quasi-classical reasoning from laws of entropy

Conditional a.g.s do not satisfy all logical laws which are valid for material implications. Neither monotony $\varphi \rightarrow \psi / \varphi \wedge \pi \rightarrow \psi$, nor contraposition $\varphi \rightarrow \psi / \neg\psi \rightarrow \neg\varphi$

nor transitivity $\varphi \to \psi, \psi \to \pi / \varphi \to \pi$ is m-valid for them. However, theorem 4.4.1 and lemma 4.4.1 tells us that in the special case where the antecedent φ of the *inferred* conditional $\varphi \Rightarrow \psi$ does *not* have a low degree-of-truth, all rules of propositional logic are m-valid. For then, we may add $\neg(\top \Rightarrow \neg\varphi)$ as a 'default premise', which say that 'φ is *practically possible*'. Let us further assume that the degree-of-truth of the conditional a.g. $\varphi \Rightarrow \psi$ is itself extremely close to 1, in other words, that this conditional a.g. is *practically necessary*. More precisely, we require that $f(\varphi \Rightarrow \psi) = f(\varphi \to \psi)/t(\varphi)$ is some decimal powers lower that $t(\varphi)$. Then it follows that all conditional a.g.s inferable from practically necessary conditional a.g.s by rules of propositional logic are themselves practically necessary, provided the antecedents of the inferred conditional a.g.s are practically possible. We call this kind of reasoning *quasi-classical reasoning*.

One area where quasi-classical reasoning is applied are physical or chemical laws involving entropy. These laws express statistical facts about the behaviour of molecules; so they are not strict laws. But due to the astronomically high number of molecules (Avogadro's number), their degree's of truth are astronomically close to 1. On the other hand, the phenomena to which these laws apply, *e.g.* phenomena of diffusion due to concentration gradients, or chemical reactions, are not astronomically rare but are practically possible – at least in our state of the universe. So it is justified to reason quasi-classically from these laws, although their logic is not really classical. For example, consider the law that a concentration gradient in a gas or fluid leads to a diffusion process. Its conditional degree of truth is astronomically close to 1 but not 1; spontaneous anti-diffusion processes with negative entropy are physically possible. Yet we may infer, by contraposition, that if in a gas or fluid no diffusion process takes places, then its matter is homogeneously distributed, *i.e.*, there is no concentration gradient. Or, we may infer by the rule of monotony that this law holds also for more specific classes of substances such as hot gases, electrolytic solutions, etc. Or we may infer, by transitivity, that if a concentration gradient between certain substances leads to a diffusion process which in turn leads to a production of heat energy, then this concentration gradient leads to a production of heat energy. All these example are instances of quasi-classical reasoning. Last but not least it follows from these considerations that the intuitive difference between 'high probability' and 'practical certainty' has a logical foundation.

That the practical certainty of quasi-classical reasoning is not guaranteed in all but in only those worlds where the antecedents of the inferred conditionals are practically possible, can be illuminated by imagining a thermodynamically equilibrated universe with maximal entropy. In such a universe, the absolute probability of concentration gradients is close to zero. Where they appear, they are caused by extremely short anti-diffusion processes. So, the law "if

concentration-gradient, then diffusion process" would not hold in this universe with high conditional probability. On the other hand, the contraposed law "if no diffusion process, then no concentration gradient" would still hold with high conditional probability. Thus, in such a universe, the contraposition from "if no diffusion process, then no concentration gradient" to "if concentration gradient, then diffusion process" would be invalid, because the absolute probability of the inferred conditional's antecedent, "concentration gradient", is close to zero in this universe.

Bibliography

E. W. Adams. Probability and the logic of conditionals. In J. Hintikka and P. Suppes, editors, *Aspects of Inductive Logic*, pages 265–316. North-Holland, Amsterdam, 1965.

E. W. Adams. The logic of 'almost all'. *Journal of Philosophical Logic*, pages 3–17, 1974.

E. W. Adams. *The Logic of Conditionals*. Reidel, Dordrecht, 1975.

E. W. Adams. A note on comparing probabilistic and modal semantics of conditionals. *Theoria*, XLII:186–194, 1977.

E. W. Adams. On the logic of high probability. *Journal of Philosophical Logic*, 15:255–279, 1986.

E. W. Adams. *A Primer of Probability Logic*. CSLI Publications, 1988.

Th. Adorno and et al. *Der Positivismusstreit in der deutschen Soziologie*. Luchterhand, Neuwied/Berlin, 1969. Edited by Maus, H. and Fürstenberg, F. and Benseler, F.

H. Albert. *Traktat über kritische Vernunft*. Mohr, Tübingen, 1968.

M. Alper. *The 'God' part of the brain. A scientific interpretation of human spirituality and God*. Sourcebook, 2008.

I. Angelelli. Frege and abstraction. *Philosophia Naturalis*, 21:453–471, 1984.

I. Angelelli. Adventures of abstraction. *Poznan Studies in the Philosophy of the Sciences and the Humanities*, 82:11–35, 2004.

I. Angelelli. Abstracci ón y pseudo-abstracción en la historia de la lógica. *Notae Philosophicae Scientiae Formalis*, 2:87–105, 2013.

G.E.M Anscombe. Causality and determination. In E. Sosa and M. Tooley, editors, *Causation*, pages 88–104. Oxford University Press, 1993.

C. Antos, R. Honzik, S.D. Friedman, and C. Ternullo. Multiverse conceptions in set theory. *Synthese*, 192(8):2463–2488, 2015.

T. Arrigoni. *What is meant by V? Reflection on universe of all sets*. Mentis Verlag, Padeborn, 2007.

T. Arrigoni. Insiemi e insiemi infiniti. Spunti dalle scienze cognizione. *Paradigmi. Rivista di critica filosofica*, pages 89–104, 2011.

T. Arrigoni and S.D. Friedman. The hyperuniverse program. *Bulletin of Symbolic Logic*, 19(1): 77–96, 2013.

S. Artemov. Explicit provability and constructive semantics. *The Bulletin of Symbolic Logic*, 7: 1–36, 2001.

S. Artemov. The logic of justification. *Review of Symbolic Logic*, 1:477–513, 2008.

S. Artemov and E. Nogina. Introducing justification into epistemic logic. *Journal of Logic and Computation*, 15:1059–1073, 2005.

G. Bacciagaluppi. A critical look at qbism. In M.C. Galavotti, D. Dieks, W.J. Gonzalez, S. Hartmann, T. Uebel, and M. Weber, editors, *New Directions in the Philosophy of Science*, pages 402–416. Springer International Switzerland, 2014.

J. Bagaria. Natural axioms of set theory and the continuum problem. In P. Hajek, L. Valdès-Villanueva, and D. Westerstahl, editors, *Logic, Methodology and Philosophy of Science. Proceedings of the Twelfth International Congress*, pages 43–63. King's College Publications, London, 2005.

A. Baltag, B. Renne, and S. Smets. The logic of justified belief, explicit knowledge and conclusive evidence. *Annals of Pure and Applied Logic*, 165:49–81, 2014.

K. Barth. Die theologie und der heutige mensch. *Zwischen der Zeiten*, 8:375–396, 1930.

K. Barth. *Fides quaerens Intellectum. Anselm Beweis der Existenz Gottes (1931)*, volume 13 of *Karl Barth-Gesamtausgabe*. Theologische Verlag Zürich, Zürich, second edition, 1931.

K. Barth. *Die kirchliche Dogmatik*. Evangelischer Verlag, Zürich, 1932–1967.

K. Barth. Philosophie und theologie. In G. Huber, editor, *Philosophie und Christliche Existenz. Festschrift für Heinrich Barth*, pages 93–106. Helbing und Lichtenhan, Basel, 1960.

K. Barth. *Der Römerbrief*, volume 47 of *Karl Barth-Gesamtausgabe*. Theologische Verlag Zürich, Zürich, second edition, 2010.

W.W. Bartley III. *The Retreat to Commitment*. Open Court, La Salle/London, second edition, 1984.

J.C. Beall. *Spandrels of Truth*. Oxford University Press, Oxford, 2009.

H. Beebee. Does anything hold the universe together? *Synthese*, 149(3):509–533, 2006.

P. Benacerraf. God, the Devil, and Gödel. *The Monist*, 51:9–32, 1967.

J.L. Bermudez and A. Cahen. Mental causation and exclusion: Why the difference-making account of causation is no help. *Humana.Mente. Journal of Philosophical Studies, special issue: Causation and Mental Causation, edited by R. Campaner and C. Gabbani*, 29:47–67, 2015.

R. Bernhardt and P. Schmidt-Leukel, editors. *Multiple religiöse Identität. Aus verschiedenen religiösen Traditionen schöpfen*, volume 5 of *Beiträge zu einer Theologie der Religionen*. Theologischer Verlag Ag, Zurich, 2008.

F. Berto. *How to Sell a Contradiction: The Logic and Metaphysics of Inconsistency*. College Publications, London, 2007.

F. Berto. The Gödel paradox and Wittgenstein's reasons. *Philosophia Mathematica*, 17: 208–219, 2009.

J. Bigelow. *The Reality of Numbers: A Physicalist's Philosophy of Mathematics*. Oxford University Press, Oxford, 1988.

A. Bindoni. Sulle definizioni per astrazione e mediante classi. *Il Bollettino di Matematica*, 11: 153–156, 1912.

M. Boden. *Computer Models of the Mind*. Cambridge University Press, Cambridge, 1988.

G. Boolos. The iterative concept of set. *Journal of Philosophy*, 68:215–232, 1971.

M. Borga, P. Freguglia, and D. Palladino, editors. *I Contributi Fondazionali della Scuola di Peano*. Franco Angeli, 1985.

R.B. Brandom. Frege's technical concepts: some recent developments. In L. Haaparanta and J. Hintikka, editors, *Frege Synthesized*, pages 253–295. Reidel, 1986.

S. Bringsjord, P. Bello, and D. Ferrucci. Creativity, the turing test, and the (better) Lovelace test. *Minds and Machines*, 11:3–27, 2001.

C. Brink and J. Heidema. A verisimilar ordering of theories phrased in a propositional language. *British Journal for the Philosophy of Science*, pages 533–549, 1987.

C.D. Broad. Determinism, indeterminism and libertarianism. In C.D. Broad, editor, *Ethics and the History of Philosophy: Selected Essays*, pages 195–217. Humanities Press, 1952.

C. Burali-Forti. Sulle classi ordinate e i numeri transfiniti. *Rendiconti del Circolo Matematico di Palermo*, VIII (169–179), 1894a.

C. Burali-Forti. *Logica Matematica*. Hoepli, 1894b.

C. Burali-Forti. Le classi finite. *Atti dell'Accademia Reale delle Scienze di Torino*, pages 34–51, 1896a.

C. Burali-Forti. Sopra un teorema del sig. G. Cantor. *Atti dell'Accademia Reale delle Scienze di Torino*, pages 229–237, 1896b.

C. Burali-Forti. Sur l'égalité et sur l'introduction des elements derives dans la science. *L'enseignement mathématique*, 1:246–261, 1899.
C. Burali-Forti. Sur le différentes methods logiques pour la definition du nombre reel. *Congrés International de Philosophie*, pages 289–307, 1901.
C. Burali-Forti. Sulle definizioni mediante "coppie". *Il Bollettino di Matematica*, VIII(237–242), 1909.
C. Burali-Forti. Gli enti astratti come enti relativi ad un campo di nozioni. *Rendiconti dell'Accademia dei Lincei*, 21(5):677–682, 1912.
C. Burali-Forti. *Logica Matematica*. Hoepli, second edition, 1919.
C. Burali-Forti. *Logica Matematica*. Edizioni della Normale, 2013. Edited by G. Lolli.
J. Burgess. The truth is never simple. *Journal of Symbolic Logic*, 51:663–681, 1987.
M. Buzzoni. Is Frankenstein's creature a machine or artificially created human life? intentionality between searle and turing. *Epistemologia*, 36:37–53, 2013.
A. Cantini. A theory of formal truth arithmetically equivalent to ID_1. *Journal of Symbolic Logic*, 55:244–259, 1990.
G. Cantor. Über unendliche, lineare punktmannifaltigkeiten 5. *Mathematische Annalen*, 21: 545–586, 1883.
G. Cantor. Beiträge zur begründung der transfiniten mengenlehre. *Mathematische Annalen*, 46, 49:215–232, 1895-1897.
I. F. Carlstrom. Truth and entailment for vague quantifiers. *Synthese*, 30:461–495, 1975.
R. Carnap. *Abriss der Logistik*. Julius Springer, 1929.
R. Carnap. *Logische syntax der Sprache*. Springer, Wien, 1934.
R. Carnap. Empiricism, semantics, and ontology. *Revue Internationale de Philosophie*, 4:20–40, 1950a.
R. Carnap. *Logical foundations of probability*. Chicago University Press, 1950b.
M. Carrara and E. Martino. Curry's paradox. a new argument for trivialism. *Logic and Philosophy of Science*, 9:199–206, 2011.
M. Carrara, S. Gaio, and E. Martino. Can Priest's dialetheism avoid trivialism? In M. Pelis and V. Puncochar, editors, *The Logica Yearbook 2011*, pages 53–64. College Publications, 2011.
U. Cassina. *Critica dei Principi della Matematics e Questioni di Logica*. Edizioni Cremonese, 1961.
S. Catania. Sulle definizioni per astrazione. *Il Bollettino di Matematica*, 10:153–156, 1911.
L.L. Cavalli-Sforza. *Genes, People, and Languages*. Farrar, Straus & Giroux, 2000.
G. Cevolani. Verisimilitude and strongly semantic information. *Ethics & Politics*, pages 159–179, 2011. http://www2.units.it/etica/.
G. Cevolani. Truth approximation via abductive belief change. *Logic Journal of the IGPL*, 21(6): 999–1016, 2013.
G. Cevolani. Strongly semantic information as information about the truth. In R. Ciuni, H. Wansing, and C. Willkommen, editors, *Recent Trends in Philosophical Logic*, pages 59–74. Springer, 2014a.
G. Cevolani. Truth approximation, belief merging, and peer disagreement. *Synthese*, 191(11): 2383–2401, 2014b.
G. Cevolani. Social epistemology, debate dynamics, and truth approximation. In U. Mäki, S. Ruphy, G. Schurz, and I. Votsis, editors, *: Recent Developments in the Philosophy of Science*, pages 57–69. Springer, 2015.
G. Cevolani. *Carnapian Truthlikeness*. Forthcoming, 2016.

G. Cevolani and F. Calandra. Approaching the truth via belief change in propositional language. In M. Suárez, M. Dorato, and M. Rèdei, editors, *EPSA Epistemology and Methodology of Science: Launch of the European Philosophy of Science Association*, pages 47–62. Springer, 2010.

G. Cevolani and V. Crupi. Subtleties of naïve reasoning: Probability, confirmation, and verisimilitude in the Linda paradox. In *Epistemology of ordinary knowledge*, pages 211–230. Cambridge Scholar, 2015.

G. Cevolani and R. Festa. Scientific change, belief dynamics and truth approximation. *La Nuova Critica*, 51–52:27–59, 2009.

G. Cevolani and R. Festa. Features of verisimilitude. Manuscript, unpublished.

G. Cevolani and G. Schurz. Probability, approximate truth, and truthlikeness: more ways out of the Preface Paradox. *Australasian Journal of Philosophy*, 95:209–225, 2017.

G. Cevolani and L. Tambolo. Progress as approximation to the truth: A defence of the verisimilitudinarian approach. *Erkenntnis*, 78(4):921–935, 2013a.

G. Cevolani and L. Tambolo. Truth may not explain predictive success, but truthlikeness does. *Studies in History and Philosophy of Science*, 44(4):590–593, 2013b.

G. Cevolani, V. Crupi, and R. Festa. The whole truth about Linda: probability, verisimilitude and a paradox of conjunction. In M. D'Agostino, G. Giorello, F. Laudisa, Pievani T., and C. Sinigaglia, editors, *SILFS. New Essays in Logic and Philosophy of Science*, pages 603–615. College Publications, London, 2010.

G. Cevolani, V. Crupi, and R. Festa. Verisimilitude and belief change for conjunctive theories. *Erkenntnis*, 75(2):183–202, 2011.

G. Cevolani, V. Crupi, and R. Festa. A verisimilitudinarian analysis of the Linda paradox. In C. Martînez Vidal, J. L. Falguera, J.M. Sagüillo, V. M. Verdejo, and M. Pereire-Fariña, editors, *VII Conference of the Spanish Society for Logic, Methodology and Philosophy of Science*, pages 366–373. USC Press, Santiago de Compostela, 2012. http://hdl.handle.net/10347/5853.

G. Cevolani, R. Festa, and T.A.F. Kuipers. Verisimilitude and belief change for nomic conjunctive theories. *Synthese*, 190(16):3307–3324, 2013.

C.S. Chihara. Priest, the liar, and Gödel. *Journal of Philosophical Logic*, 13(2):117–124, 1984.

M. Cipolla. *Analisi Algebrica e Introduzione al Calcolo Infinitesimale*. Capozzi, 1914.

R.K. Clarke. *Libertarian Accounts of Free Will*. Oxford University Press, 2003.

D. Cockburn. Tense and emotion. In R. LePoidevin, editor, *Questions of Time and Tense*, pages 77–91. Clarendon Press, 1998.

K. Coleman and E.O. Wiley. New defense of the species-as-individuals hypothesis. *Philosophy of Science*, 68(4):498–517, 2001.

M. Colyvan. Vagueness and truth. In H. Dyke, editor, *From Truth to Reality: New Essays in Logic and Metaphysics*, pages 29–40. Routledge, 2009.

K. Corcoran. The trouble with Searle's biological naturalism. *Erkenntnis*, 55:307–324, 2001.

A. Corradini. Mental causation for mind-body dualists. *Humana.Mente. Journal of Philosophical Studies, special issue: Causation and Mental Causation, edited by R. Campaner and C. Gabbani*, 29:91–124, 2015.

F. Correia and B. Schnieder, editors. *Metaphysical Grounding: Understanding the Structure of Reality*. Cambridge University Press, Cambridge, 2012.

L. Couturat. *Les Principes des Mathématiques*. Alcan, 1905.

J. Crane. On the metaphysics of species. *Philosophy of Science*, 71:156–173, 2004.

T. Crane. *The Mechanical Mind: A Philosophical Introduction to Minds, Machines, and Mental Representation*. Penguin, London, second edition, 2003.

T. Crane. Causation and determinable properties: On the efficacy of colour, shape, and size. In H. Jakob and J. Kallestrup, editors, *Being Reduced: New Essays on Reduction, Explanation, and Causation*, pages 176–195. Oxford University Press, Oxford, 2008.

C. Darwin. *The Descent of Man, and Selection in Relation to Sex*. John Murray, London, 1871.

A.P. Dawid and M.C. Galavotti. De Finetti's subjectivism, objective probability, and the empirical validation of probability assessments. In M.C. Galavotti, editor, *Bruno de Finetti, Radical Probabilist*, pages 97–114. College Publications, 2009.

E. De Amicis. Dipendenza fra alcune proprietá notevoli delle relazioni fra enti di un medesimo sistema. *Rivista di Matematics*, 1:113–127, 1892.

B. de Finetti. La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7:1–68, 1937. English edition: "Foresight: its Logical Laws, its Subjective Sources", in H. Kyburg and H. Smokler (eds.) *Studies in Subjective Probability*, New York-London-Sydney, Wiley, pp. 95-158.

B. de Finetti. Does it make sense to speak of 'good probability appraisers'? In I.J. Good, A.J. Mayne, and J.M. Smith, editors, *The Scientist Speculates. An Anthology of Partly-Baked Ideas*, pages 357–364. Basic Books, 1962a.

B. de Finetti. Obiettività e oggettività: critica a un miraggio. *La Rivista Trimestrale*, 1:343–367, 1962b.

B. de Finetti. *Teoria delle probabilità*. Einaudi, 1970. English edition: *Theory of Probability*, New York: Wiley, 1975.

B. de Finetti. Bayesianism: Its unifying role for both the foundations and the applications of statistics. *Bulletin of the International Statistical Institute*, pages pp. 349–368, 1973.

B. de Finetti. The value of studying subjective evaluations of probability. In C.A.S. Staël von Holstein, editor, *The Concept of Probability in Psychological Experiments*, pages 1–14. Reidel, 1974a.

B. de Finetti. The true subjective probability problem. In C.A.-S. Staël von Holstein, editor, *The Concept of Probability in Psychological Experiments*, pages 15–23. Reidel, 1974b.

B. de Finetti. *Filosofia della Probabilità*. Il Saggiatore, Milano, 1995. English edition: *Philosophical Lectures on Probability*, ed. by A. Mura, Dordrecht, Springer, 2008.

W. Dean and H. Kurokawa. From the knowability paradox to existence of proofs. *Synthese*, 176: 177–225, 2010.

D. Dennett. *The Intentional Stance*. MIT Press, Cambridge, 1987.

D. Dennett. *Darwin's dangerous idea. Evolution and the meanings of life*. Simon and Shuster, New York, 1995.

K. Devlin. *Constructibility*. Springer, Berlin-New York, 1984.

W. Dilthey. *Einleitung in die Geisteswissenschaften*. Duncker & Humblot, Leipzig, 1883.

T. Dobzhansky. A critique of the species concept in biology. *Philosophy of Science*, 2(3): 344–355, 1935.

T. Dobzhansky. *Genetics and the Origin of Species*. Columbia University Press, New York, 1937.

W. Dubislav. *Die Definition*. Meiner, third edition, 1931.

H. N. Duc. Reasoning about rational, but not logically omniscient, agents. *Journal of Logic and Computation*, 7:633–648, 1997.

M. Dummett. Wittgenstein's philosophy of mathematics. *The Philosophical Review*, 68(3): 324–348, 1959.

M. Dummett. *Frege. Philosophy of mathematics*. Harvard University Press, Cambridge MA, 1991.

P. Egré. The knower paradox in the light of provability interpretations of modal logic. *Journal of Logic, Language and Information*, 14:13–48, 2005.

N. Eldredge and J. Cracraft. *Phylogenetic Patterns and the Evolutionary Process*. Columbia University Press, New York, 1980.

M. Ereshefsky. *The Poverty of the Linnaean Hierarchy, A Philosophical Study of Biological Taxonomy*. Cambridge University Press, Cambridge, second edition, 2003.

G. Eriksson. Linnaeus the botanist. In T. Frängsmyr, editor, *Linnaeus the Man and His Work*, pages 63–109. University of California Press, Berkeley, 1983.

G. Erion. The cartesian test for automatism. *Minds and Machines*, 11:29–39, 2001.

M. Esfeld. Mental causation and the metaphysics of causation. *Erkenntnis*, 67(2):207–220, 2007.

S. Feferman. Penrose's Gödelian argument. a review of *Shadows of the Mind* by Roger Penrose. *Psyche*, 2(7), 1995.

S. Feferman. Are there absolutely unsolvable problems? Gödel's dichotomy. *Philosophia Mathematica*, 14:134–152, 2006.

S. Feferman. Gödel, Nagel, minds, and machines. *The Journal of Philosophy*, 106(4):201–219, 2009.

R. Festa. Theory of similarity, similarity of theories, and verisimilitude. In T.A.F. Kuipers, editor, *What is Closer-to-the-truth*, Poznan Studies in the Philosphy of Sciences and the Humanities, pages 145–176. Rodopi, Amsterdam, 1987.

R. Festa. Verisimilitude, cross classification, and prediction logic. Approaching the statistical truth by falsified qualitative theories. *Mind and Society*, 6:37–62, 2007a.

R. Festa. The qualitative and statistical verisimilitude of qualitative theories. *La Nuova Critica*, 47–48:91–114, 2007b.

R. Festa. Verisimilitude, qualitative theories, and statistical inferences. In S. Pihlström, P. Raatikainen, and M. Sintonen, editors, *Approaching the Truth. Essays in Honour of Ilkka Niiniluoto*, pages 143–178. College Publications, London, 2007c.

R. Festa. The statistical verisimilitude of qualitative theories. *La Nuova Critica*, 53–54:51–78, 2009.

R. Festa. Bayesian inductive logic, verisimilitude, and statistics. In D.M. Gabbay and J. Woods, editors, *Handbook of The Philosophy of Science: Philosophy of Statistics*, pages 473–490. North Holland, San Diego, 2011.

R. Festa. On the verisimilitude of tendency hypotheses. In D. Dieks, W.J. Gonzalez, S. Hartmann, M. Stöltzner, and M. Weber, editors, *Probabilities, Laws, and Structures. The Philosophy of Science in a European Perspective*, volume 3, pages 43–55. Springer, Dordrecht, 2012.

H. Field. *Science without Numbers*. Princeton University Press, Princeton, 1980.

H. Field. *Realism, Mathematics and Modality*. Basil Blackwell, Oxford, 1989.

K. Fine. Vagueness, truth and logic. *Synthese*, 30:265–300, 1975.

K. Fine. The question of realism. *Philosophers Imprint*, 1(1):1–30, 2001.

K. Fine. A guide to ground. In F. Correia and B. Schnieder, editors, *Metaphysical Grounding: Understanding the Structure of Reality.*, pages 37–80. Cambridge University Press, Cambridge, 2012.

M. Fischer and J. Stern. Paradoxes of interaction? *Journal of Philosophical Logic*, 44:287–308, 2015.

M. Fischer, V. Halbach, J. Krin, and J. Stern. Axiomatizing semantic theories of truth? *Review of Symbolic Logic*, 8:257–278, 2015.

E. Fisher, J.J. von Schomberg, and R. von Schomberg, editors. *Implementing the Precautionary Principle: Perspectives and Prospects*. Edward Elgar, Chltenham UK and Nothampton (MA), 2006.

M. Fitting. The logic of proofs, semantically. *Annals of Pure and Applied Logic*, 132:1–25, 2005.

M. Fitting. The logic of proofs, semantically. *Annals of Pure and Applied Logic*, 152:67–83, 2008.

O. Flanagan. *The Science of the Mind*. MIT Press, Cambdrige (Mass), second edition, 1991.

J. Fodor. Searle on what only brains can do. *Behavioral and Brain Sciences*, 3:431–432, 1980.

J. Fodor. *In Critical Condition: Polemical Essays on Cognitive Science and the Philosophy of Mind*. MIT Press, Cambdrige (Mass), 1998.

K.R. Foster, P. Vecchia, and M.H. Repacholi. Science and the precautionary principle. *Science*, 288(5468):979–981, 2000.

D.A. Freedman. Some issues in the foundations of statistics. *Foundations of Science*, 1:19–39, 1995.

D.A. Freedman and P.B. Stark. What is the chance of an earthquake? Technical Report 611, September 2001, revised January 2003; online at: tr-611.tex; 28/01/2003, 2003.

G. Frege. *Grundlagen der Arithmetik*. WBG, 1884.

G. Frege. *Kleine Schriften*. WBG, 1967.

G. Frege. *Wissenschaftlicher Briefwechsel*. Meiner, 1976. Edited by G. Gabriel, H. Hermes, F. Kambartel, C. Thiel, A. Veraart.

G. Frege. *Nachgelassene Schriften*. Meiner, 1983. Edited by H. Hermes, F. Kambartel, F. Kaulbach.

P. Freguglia. Definizione per astrazione e numeri cardinali. *Cultura e Scuola*, 81:199–208, 1982.

R.M. French. Subcognition and the limits of the Turing test. *Mind*, 99:53–65, 1990.

R.M. French. Refocusing the debate on the Turing test: A reply to Jacquette. *Behavioral and Brain Sciences*, 23(1):61–62, 1995.

R.M. French. Peeking behind the screen: the unsuspected power of the standard Turing test. *Journal of Experimental and Theoretical Artificial Intelligence*, 12:331–340, 2000.

H. Gaifman. What Gödel's incompleteness result does and does not show. *The Journal of Philosophy*, 97:462–470, 2000.

M.C. Galavotti. F.P. Ramsey and the notion of 'chance'. In J. Hintikka and K. Puhl, editors, *The British Tradition in the 20th Century Philosophy. Proceedings of the 17th International Wittgenstein Symposium*, pages pp. 330–340. Hölder-Pichler-Tempsky, 1995.

M.C. Galavotti. Some remarks on objective chance (F.P. Ramsey, K.R. Popper and N.R. Campbell). In M.L. Dalla Chiara, R. Giuntini, and F. Laudisa, editors, *Language, Quantum, Music*, chapter 73–82. Kluwer, 1999.

M.C. Galavotti. Subjectivism, objectivism and objectivity in Bruno de Finetti's Bayesianism. In D. Corfield and J. Williamson, editors, *Foundations of Bayesianism*, pages 173–186. Kluwer, 2001.

M.C. Galavotti. *Philosophical Introduction to Probability*. CSLI, 2005.

M.C. Galavotti. Probability: One or many? In B. Löwe, E. Pacuit, and J.-W. Romeijn, editors, *Foundations of the Formal Sciences VI: Reasoning about Probabilities and Probabilistic Reasoning*, pages 153–170. College Publications, 2009.

M.C. Galavotti. Probability, statistics, and law. In D. Dieks, S. Hartmann, M. Stoeltzner, M. Weber, and W.J. Gonzalez, editors, *Probability, Laws, and Structures*, pages 401–412. Springer, 2012.

R. M. Gale. Time, temporality, and paradox. In R. M. Gale, editor, *The Blackwell Guide to Metaphysics*, pages 66–86. Blackwell, 2002.

S. Galvan. Il concetto di verità in A. Tarski. *Verifiche*, 2:3–66, 1973.

S. Galvan. *Teoria formale dei numeri naturali*. FrancoAngeli, Milano, 1983.

S. Galvan. *Introduzione ai teoremi di incompletezza*. FrancoAngeli, Milano, 1992.

S. Galvan. Gödel e il modello computazionale della mente. *Rivista di Filosofia Neo-Scolastica*, 96:145–174, 2004.

R. Gelman. First principles organize attention to and learning about relevant data: Number and the animate-Inanimate distinction as examples. *Cognitive Science*, 14:79–106, 1990.

K. Gemes. Verisimilitude and content. *Synthese*, 154(2):293–306, 2007.

G. Gentzen. Untersuchungen über das logische schliessen. *Math. Zeits.*, 39:405–431, 1935.

G. Gergely and G. Csibra. Teleological reasoning in infancy: the naïve theory of rational action. *Trends in Cognitive Sciences*, 7(7):287–292, 2003.

M. Ghiselin. On psychologism in the logic of taxonomic controversies. *Systematic Zoology*, 15: 207–215, 1966.

M. Ghiselin. A radical solution to the species problem. *Systematic Zoology*, 23(4):536–544, 1974.

M. Ghiselin. Species concepts, individuality, and objectivity. *Biology and Philosophy*, 2: 127–143, 1987.

A. Giordani. A new semantics for systems of logic of essence. *Studia Logica*, 102:411–440, 2013.

J-Y. Girard. Une extension de l'interprétation de gödel à l'analyse et son application à l'élimination des coupures. In J.B. Fenstad, editor, *Proceedings of 3nd Scandinavian Logic Symposium*, pages 63–92. North-Holland, 1971.

K. Gödel. Russell's mathematical logic. In P. A. Schilpp, editor, *The Philosophy of Bertrand Russell*. Evanstone, 1944.

K. Gödel. What is Cantor's continuum problem? In P. Benacerraf and H. Putnam, editors, *Philosophy of Mathematics: Selected Readings*, pages 470–485. Cambridge University Press, Cambridge, 1983.

K. Gödel. *Collected Works. Vol. I. Publications 1929-1936*. Oxford University Press, Oxford, 1986. Edited by S. Feferman *et al.*

K. Gödel. *Collected Works. vol. III. Unpublished Essays and Lectures*. Oxford University Press, Oxford, 1995. Edited by S. Feferman *et al.*

A. Goldman. *Knowledge in a Social World*. Clarendon Press, Oxford, 1999.

C. Gollier, J. Bruno, and N. Treich. Scientific progress and irreversibility: An economic interpretation of the 'precautionary principle'. *Journal of Public Economics*, 75:229–253, 2000.

I.J. Good. *The Estimation of Probabilities, An Essay on Modern Bayesian Methods*. MIT Press, 1965.

I.J. Good, A.J. Mayne, and J.M. Smith, editors. *The Scientist Speculates. An Anthology of Partly-Baked Ideas*. Basic Books, 1962.

I. Grattan-Guinness. *The Search for Mathematical Roots 1870-1940. Logics, set theories, and the foundations of mathematics from Cantor through Russell and Gödel*. Princeton University Press, 2000.

A. Gupta. Truth and paradox. *Journal of Philosophical Logic*, 11:1–60, 1982.

T. Hailperin. Probability semantics for quantifier logic. *Journal of Philosophical Logic*, 29: 207–239, 2000.

V. Halbach. Disquotational truth and analyticity. *Journal of Symbolic Logic*, 66:1959–1973, 2001.
V. Halbach. How not to state the T-sentences. *Analysis*, 66:276–280, 2006.
V. Halbach. On a side effect of solving Fitch's paradox by typing knowledge. *Analysis*, 68: 114–120, 2008.
V. Halbach. *Axiomatic theories of truth*. Cambridge University Press, Cambridge, 2014.
V. Halbach and P. Welch. Necessities and necessary truths: a prolegomenon to the use of modal logic in the analysis of intensional notions. *Mind*, 118:71–100, 2009.
V. Halbach, H. Leitgeb, and P. Welch. Possible world semantics for modal notions conceived as predicates. *Journal of Philosophical Logic*, 32:179–233, 2003.
J. Haldane. Naturalism and intentionality. *Inquiry*, 32:305–322, 1989.
N. Hall. Two concepts of causation. In J. Collins, N. Hall, and L. Paul, editors, *Causation and Counterfactuals*, pages 225–276. MIT Press, Cambdrige (Mass), 2004.
F Harary and P. Per Hage. *Structural Models in Anthropology*. Cambridge University Press, Cambridge, 1983.
A. von Harnack. *Das Wesen des Christentums*. J.C. Hinrichs, Leipzig, 1900.
S. Harnad. Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines*, 1:43–54, 1991.
J. Haugeland. Syntax, semantics, physics. In J. Preston and M. Bishop, editors, *Views into the Chinese Room: Essays on Searle and Artificial Intelligence*, pages 379–392. Clarendon Press, Oxford, 2002.
F. Hausdorff. *Grundzüge der Mengenlehre*. Veit, 1914.
K. Hauser. Objectivity over objects. A case study in theory formation. *Synthese*, 128:245–285, 2001.
K. Hauser. Gödel's program revisited. The turn to phenomenology. *Bulletin of Symbolic Logic*, 12(4):529–560, 2006.
L. Hauser. Nixon' goes to China. In J. Preston and M. Bishop, editors, *Views into the Chinese Room: Essays on Searle and Artificial Intelligence*, pages 123–143. Clarendon Press, Oxford, 2002.
J. Hawthorne. On the logic of nonmmonotonic conditionals and conditional probabili- ties: Predicate logic. *Journal of Philosophical Logic*, 27:1–34, 1998.
R. Held and A. Hein. Movement-produced stimulation in the development of visually guided behavior. *Journal of Comparative and Physiological Psychology*, 56(5):872–876, 1963.
H. Helmholtz. Zählen und messen, erkenntnistheoretisch betrachtet. In *Philosophische Aufsätze, Eduard Zeller zu seinem fünfzigjährigen Doctorjubiläum gewidmet*, pages 17–52. Fues' Verlag, 1887.
A. Hendry. Questioning species realities. *Conservation Genetics*, 1:67–76, 2000.
W. Henning. *Phylogenetic Systematics*. Press of Chicago University, Chicago, 2001.
J. Hey. *Genes, Categories, and Species. The Evolutionary and Cognitive Causes of the Species Problem*. Oxford University Press, Oxford, 2001.
J. Hick. *An Interpretation of Religion. Human Responses to the Transcendent*. Macmillan, Basingstoke, second edition, 1992.
D. Hilbert. Die logischen grundlagen der mathematik. *Mathematische Annalen*, 88:151–165, 1923.
D. Hilbert. Über das unendliche. *Mathematische Annalen*, 95:161–190, 1926.
D. Hilbert and P. Bernays. *Grundlagen der mathematik*, volume I. Springer, 1934.
K. Holsinger. The nature of biological species. *Philosophy of Science*, 51(2):293–307, 1984.

L. Horsten and H. Leitgeb. No future. *Journal of Philosophical Logic*, 30:259–265, 2001.
W.A. Howard. The formulae-as-types notion of construction. In Hindley J.R. and Seldin J.P., editors, *To H.B. Curry: Essays on combinatory logic, Lambda Calculus and Formalism*, pages 479–490. Academic Press, 1980.
G. E. Hughes and M. J. Cresswell. *A new introduction to modal logic*. Routledge, London, 1996.
D. Hull. The effect of essentialism on taxonomy – two thousand years of stasis (i). *The British Journal for the Philosophy of Science*, 15(60):314–326, 1965a.
D. Hull. The effect of essentialism on taxonomy – two thousand years of stasis (ii). *The British Journal for the Philosophy of Science*, 15(60):314–326, 1965b.
D. Hull. Are species really individuals? *Systematic Zoology*, 25(2):174–191, 1976.
D. Hull. A matter of individuality. *Philosophy of Science*, 45(3):335–360, 1978.
D. Hume. *An Enquiry Concerning Human Understanding*. Oxford University Press, Oxford, 1999. Edited by T. L. Beauchamp.
D. Jacquette. Adventures in the chinese room. *Philosophy and Phenomenological Research*, 49: 605–623, 1989.
I. Jané. The role of the absolute infinite in Cantor's conception of set. *Erkenntnis*, 42:375–402, 1995.
I. Jané. The iterative concept of set from a Cantorian perspective. In P. Hajek, L. Valdès-Villanueva, and D. Westerstahl, editors, *Logic, Methodology and Philosophy of Science. Proceedings of the Twelfth International Congress*, pages 373–393. King's College Publications, 2005.
T. Jech. *Set Theory*. Springer, 2003.
A. Kanamori. *The Higher Infinity*. Springer, Berlin, 1994.
G. Keil. *Handeln und Verursachen*. Klostermann, 2000.
J. Kim. *Physicalism, or Something Near Enough*. Princeton University Press, Princeton, 2005.
S. Kripke. Outline of a theory of truth. *Journal of Philosophy*, 72:690–716, 1975.
V. Kuhlmeier, P. Bloom, and K. Wynn. Do 5-month-old infants see humans as material objects? *Cognition*, 94:95–103, 2004.
T.A.F. Kuipers. Approaching descriptive and theoretical truth. *Erkenntnis*, 18:343–378, 1982.
T.A.F. Kuipers, editor. *What is Closer-to-the-truth?* Rodopi, Amsterdam, 1987a.
T.A.F. Kuipers. A structuralist approach to truthlikeness. In *What is Closer-to-the-truth?*, pages 79–99. Rodopi, Amsterdam, 1987b.
T.A.F. Kuipers. Basic and refined nomic truth approximation by evidence-guided belief revision in AGM-terms. *Erkenntnis*, 75(2):223–236, 2011.
T.A.F. Kuipers. Models, postulates, and generalized nomic truth approximation. *Synthese*, 2015. DOi: 10.1007/s11229-015-0916-9.
K. Kunen. *Set Theory. An Introduction to Independence Proofs*. North-Holland, Amsterdam, 1980.
W. Kunz. *Do Species Exist? Principles of Taxonomic Classification*. Wiley-Blackwell, Weinheim, 2012.
F. v. Kutschera. *Gottlob Frege*. W. de Gruyter, 1989.
F. v. Kutschera. Concepts of a set. In R. Stuhlmann-Laeisz A. Newen, U. Nortmann, editor, *Building on Frege – New Essays on Sense, Content, and Concept*, pages 319–327. Stanford CSLI, 2001.
A.B. Laksana. Multiple religious belonging or complex identity? an asian way of being religious. In *The Oxford Handbook of Christianity in Asia*, pages 493–509. Oxford University Press, 2014.

J. LaPorte. *Natural Kinds and Conceptual Change*. Cambridge University Press, Cambridge, 2004.
J. Larson. The species concept of Linnaeus. *Isis*, 59(3):291–299, 1968.
L. Laudan. Is reasonable doubt reasonable? *Legal Theory*, 9:295–331, 2003.
L. Laudan. *Truth, Error, and Criminal Law*. Cambridge University Press, 2006.
D. Lehmann and M. Magidor. What does a conditional knowledge base entail? *Arificial Intelligence*, 55:1–60, 1992.
W. Lenzen. Leibniz und die entwicklung der modernen logik. In *Leibniz Werk und Wirkung – Akten des IV. Internationalen Leibniz Kongress*, pages 418–425, Hannover, 1983.
W. Lenzen. 'Non est' non est, 'est non' – Zu Leibnizens theorie der negation. *Studia Leibnitana*, 18:1–37, 1986.
W. Lenzen. Guilielmi pacidii non plus ultra – eine rekonstruktion von Leibniz' plus-minus-kalkül. *Philosophiegeschichte und Logische Analyse*, 3:71–118, 2000.
W. Lenzen. Logical criteria for individual(concept)s. In M. Carrara, A.M. Nunziante, and G. Tomasi, editors, *Individuals, Minds, and Bodies: Themes from Leibniz*, pages 87–107. Steiner Verlag, Stuttgart, 2004.
G.E. Lessing. *Die Erziehung des Menschengeschlechts (1780)*, volume 10 of *Werke und Briefe in 12 Bänden*, pages 73–99. Suhrkamp, Frankfurt a.M., 2001.
D. Lewis. *Counterfactuals*. Basil Blackwell, Oxford, 1973.
D. Lewis. *Papers in Metaphysics and Epistemology*. Cambridge University Press, 1999.
C. List and P. Menzies. Nonreductive physicalism and the limits of the exclusion principle. *Journal of Philosophy*, 1006:475–502, 2009.
W. Loeffler. *Einführung in die religionsphilosophie*. Wissenschaftliche Buchgesellschaft, Darmstadt, 2006.
G. Lolli. Cesare Burali-Forti (1861–1931) e la logica matematica. In G. Lolli, editor, *Logica Matematica*, pages vii–lxiii. Edizioni della Normale, 2013.
A. Lovejoy. Buffon and the problem of species. In B. Glass, O. Temkin, and W.L. Straus, editors, *Forerunners of Darwin*, pages 84–113. John Hopkins University Press, Baltimore, 1968.
E.J. Lowe. *Personal agency. The Metaphysics of Mind and Action*. Oxford University Press, 2008.
K. Löwith. *Meaning in History. The Theological Implications of the Philosophy of History*. University of Chicago Press, Chicago-London, 1949.
J.R. Lucas. Minds, machines and Gödel. *Philosophy*, 36:112–127, 1961.
R. D. Luce. Semiorders and a theory of utility discrimination. *Econometrica*, 24:178–191, 1956.
E. Maccaferri. Le definizioni per astrazione e la classe di Russell. *Rendiconti del Circolo Matematico di Palermo*, pages 165–171, 1913.
V. Mago. Teoria degli ordini. *Memorie della Reale Accademia delle scienze di Torino*, 64(2):1–25, 1913.
D. Makinson. General patterns in nonmonotonic reasoning. In D. Gabbay, editor, *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 3, pages 35–110. Clarendon Press, Oxford, 1994.
P. Mancosu. Grundlagen, §64: Frege's discussion of definitions by abstraction in historical context. *History and Philosophy of Logic*, 36:62–89, 2015a.
P. Mancosu. In good company? On Hume's principle and the assignment of numbers to infinite concepts. *The Review of Symbolic Logic*, 8(2):370–410, 2015b.
D.A. Martin. Mathematical evidence. In H.G. Dales and G. Oliveri, editors, *Truth in Mathematics*, pages 215–231. Clarendon Press, 1998.

D.A. Martin and J. Steel. A proof of projective determinacy. *Journal of the American Mathematical Society*, 2:71–125, 1989.
E. Mayr. *Systematic and the Origin of Species*. Columbia University Press, New York, 1942.
E. Mayr. The species concept: Semantics versus semantics. *Evolution*, 3:371–372, 1949.
E. Mayr. *Populations, Species, and Evolution*. Harvard University Press, Cambdrige (Mass), 1970.
S. McCall. Can a Turing machine know that the gödel sentence is true? *The Journal of Philosophy*, 96(10):525–532, 1999.
V. McGee. *Truth, Vagueness and Paradox. An essay on the logic of truth*. Hackett, Indianapolis, 1991.
C. McGinn. *The Mysterious Flame: Conscious Minds in a Material World*. Basic Books, New York, 1991.
C. McGinn. *Consciousness and Its Objects*. Clarendon Press, Oxford, 2004.
U. Meixner. *Axiomatic Formal Ontology*. Kluwer, 1997.
U. Meixner. *Theorie der Kausalität. Ein Leitfaden zum Kausalbegriff in zwei Teilen*. Mentis, 2001.
U. Meixner. *The Theory of Ontic Modalities*. ontos, 2006.
A. Melnyk. Searle's abstract argument against strong AI. *Synthese*, 108:391–419, 1996.
P. Menzies. The exclusion problem, the determination relation, and contrastive causation. In J. Hohwy, editor, *Being reduced. New essays on reduction, explanation, and causation.*, pages 196–217. Oxford University Press, 2008.
P. Menzies. Mental causation in the physical world. In S.C. Gibb, R. Ingthorsson, and E.J. Lowe, editors, *Mental Causation and Ontology*, pages 58–87. Oxford University Press, Oxford, 2013.
P. Menzies. The causal closure argument is no threat to non-reductive physicalism. *Humana.Mente. Journal of Philosophical Studies, special issue: Causation and Mental Causation, edited by R. Campaner and C. Gabbani*, 29:21–46, 2015.
D. Miller. Popper's qualitative theory of verisimilitude. *The British Journal for the Philosophy of Science*, 25(2):166–177, 1974.
D. Miller. The distance between constituents. *Synthese*, 38(2):197–212, 1978.
D. Miller. *Critical rationalism: a restatement and defence*. Open Court, Chicago, 1994.
D. Miller. *Out Of Error: Further Essays On Critical Rationalism*. Ashgate, London, 2006.
G. Molnar. *Powers. A Study in Metaphysics*. Clarendon Press, Oxford, 2006. Edited by S. Mumford.
R. Montague. Universal grammar. *Theoria*, XXXVI:373–398, 1970.
R. Montague. Syntactical treatments of modality with corollaries on reflexion principles and finite axiomatizability. In R. Thomason, editor, *Formal Philosophy.Selected papers of Richard Montague*. Yale University Press, New Haven, 1974.
C. Mortensen. A theorem of verisimilitude. *BSL*, 7:34–43, 1978.
C. Mortensen. Relevance and verisimilitude. *Synthese*, 55(3):353–364, 1983.
Y. Moschovakis. *Elementary induction on abstract structures*. North-Holland, Amsterdam, 1974.
Y. Moschovakis. *Descriptive set theory*. North-Holland, Amsterdam, 1980.
P.L. Mott. Verisimilitude by means of short theorems. *Synthese*, 38(2):247–273, 1978.
K. Müller and G. Krönert. *Leben und Werk von Gottfried Wilhelm Leibniz – Eine Chronik*. Vittorio Klostermann, Frankfurt a.M., 1969.
S. Mumford. Filled in space. In M. Kistler and B. Gnassounou, editors, *Dispositions and Causal Powers*, pages 67–80. Ashgate, 2007.
S. Mumford. Passing powers around. *The Monist*, 92(1):94–111, 2009.

S. Mumford and R.L. Anjum. *Getting Causes from Powers*. Oxford University Press, 2011.
J. Murzi and M. Carrara. Paradox and logical revision. A short introduction. *Topoi*, 34:7–14, 2015.
A. Natucci. *Il Concetto di Numero e le sue Estensioni*. Bocca, 1923.
G. Nelson and N. Plamick. *Systematics and Biogeography: Cladistics and Vicariance*. Columbia University Press, New York, 1981.
M. W. Nicholson. Abusing Wittgenstein: The misuse of the concept of language games in contemporary theology. *JETS*, 39:359–369, 1996.
I. Niiniluoto. *Truthlikeness*. Reidel, Dordrecht, 1987.
I. Niiniluoto. Verisimilitude: the third period. *British Journal for the Philosophy of Science*, 49 (1):1–29, 1998.
I. Niiniluoto. Scientific progress as increasing verisimilitude. *Studies in History and Philosophy of Science*, 46:73–77, 2014.
I. Niiniluoto. Optimistic realism about scientific progress. *Synthese*, pages 1–19, 2015.
A. Noë. *Action in Perception*. MIT Press, 2004.
R. Northcott. Verisimilitude: A causal approach. *Synthese*, 190(9):1471–1488, 2013.
K. Obermeier. Wittgenstein on language and artificial intelligence: The chinese-room thought experiment rivisited. *Synthese*, 56:339–349, 1983.
T. O'Connor. Libertarian views: Dualist and agent-causal theories. In R. Kane, editor, *The Oxford Handbook of Free Will*, pages 337–355. Oxford University Press, 2001.
G. Oddie. *Likeness to truth*. Reidel, Dordrecht, 1986.
G. Oddie. Verisimilitude and the convexity of propositions. In T.A.F. Kuipers, editor, *'What is Closer-to-the-Truth?*, pages 197–216. Rodopi, Amsterdam, 1987.
G. Oddie. The content, consequence and likeness approaches to verisimilitude: compatibility, trivialization, and underdetermination. *Synthese*, 190:1647–1687, 2013.
G. Oddie. Truthlikeness. In E.N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. CSLI Publications, 2014.
S. Okasha. Darwinian metaphysics: Species and the question of essentialism. *Synthese*, 131: 191–213, 2002.
A. Padoa. Dell'astrazione matematica. *Questioni Filosofiche*, pages 91–104, 1908.
W. Pannenberg. *Wissenschaftstheorie und Theologie*. Suhrkamp, Frankfurt a.M., 1973.
G. Peano. *Calcolo geometrico secondo l'Audehnungslehre di H. Grassmann*. Bocca, 1888.
G. Peano. *Notations de Logique Mathématique. Introduction au Formulaire de mathématique*. Guadagnini, 1894.
G. Peano. *Formulaire de Mathématiques, publié par la Revue de Mathématiques. t. II, n. 3. Logique mathématique. Arithmétique. Limites. Nombres complexes. Vecteurs. Dérivés. Intégrales*. Bocca, 1899a.
G. Peano. Sui numeri irrazionali. *Rivista di Matematica*, 6:126–140, 1899b.
G. Peano. Formules de logique mathématique. *Rivista di Matematica*, 7(1–41), 1900.
G. Peano. Les definitions mathématiques. *Congrés International de Philosophie*, pages 279–288, 1901a.
G. Peano. *Formulaire de Mathématiques, t. III*. Bocca, 1901b.
G. Peano. Dizionario di matematica. Parte I, logica matematica. *Rivista di Matematica*, 7: 160–172, 1901c.
G. Peano. *Le definizioni in matematica*. Institut d'Estudis Catalans, Barcelona, 1911.
G. Peano. Le definizioni per astrazioni. *Bollettino della Mathesis*, VII:106–120, 1915.

P.C. Phan. Multiple religious belonging: Opportunities and challenges for theology and church. *Theological Studies*, 64:498–522, 2003.
W. Pohlers. *Proof Theory. First steps into impredicativity*. Springer, Berlin-New York, 2009.
K.R. Popper. *The Open Society and Its Enemies*. Routledge, London, 1945.
K.R. Popper. *The Poverty of Historicism*. Routledge, London, 1957.
K.R. Popper. On the sources of knowledge and of ignorance (1960). In *Conjecture and Refutations. The Growth of Scientific Knowledge*, pages 3–30. Routledge, London, 1963a.
K.R. Popper. *Conjectures and Refutations: the Growth of Scientific Knowledge*. Routledge and Kegan Paul, London, 1963b.
D. Poulin-Dubois, A. Lepage, and D. Ferland. Infants' concept of animacy. *Cognitive Development*, 11(1):19–36, 1996.
B. Preston. The ontological argument against the mind-machine hypothesis. *Philosophical Studies*, 80:131–157, 1995.
G. Priest. The logic of paradox. *Journal of Philosophical Logic*, 8:219–241, 1979.
G. Priest. Is arithmetic consistent? *Mind*, 103:337–349, 1994.
G. Priest. *Beyond the Limits of Thought*. Oxford University Press, Oxford, 2001.
G. Priest. Paraconsistent logic. In D.M. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 6, pages 287–393. Kluwer Academic Publisher, 2002.
G. Priest. *Doubt Truth To Be A Liar*. Oxford University Press, Oxford, 2006a.
G. Priest. *In Contradiction*. Oxford University Press, Oxford, second edition, 2006b.
H. Putnam. Minds and machines. In S. Hook, editor, *Dimensions of Mind: A Symposium*, pages 148–179. University of New York Press, New York, 1960.
H. Putnam. *The Project of Artificial Intelligence*, pages 1–18. Harvard University Press, Cambdrige MA, 1992.
W. v. O. Quine. *Word and Object*. M.I.T. Press, 1960.
P. Raatikainen. Mccall's Godelian argument is invalid. *Facta Philosophica*, 4:167–169, 2002.
P. Raatikainen. Hilbert's program rivisited. *Synthese*, 137:157–177, 2003.
P. Raatikainen. On the philosophical relevance of gödel's incompleteness theorems. *Revue Internationale de Philosophie*, 59(234):513–534, 2005.
P. Raatikainen. Gödel's incompleteness theorems. In E. Zalta, editor, *Stanford Encyclopedia of Philosophy*. Stanford CSLI, 2015.
F.P. Ramsey. *The Foundations of Mathematics and Other Logical Essays*. Routledge and Kegan Paul, 1931. Edited by R.B. Braithwaite.
F.P. Ramsey. *Philosophical Essays*. Cambridge University Press, 1990. Edited by H. Mellor.
F.P. Ramsey. *Notes on Philosophy, Probability and Mathematics*. Bibliopolis, 1991. Edited by M. C. Galavotti.
M. C. Rea. Four-dimensionalism. In M. Loux a. D. Zimmerman, editor, *The Oxford Handbook of Metaphysics*, pages 246–280. Oxford University Press, Oxford, 2003.
E. Reck. Dedekind's structuralism. *Synthese*, 137:369–419, 2003.
E. Reck and M. Price. Structures and structuralism in contemporary philosophy of mathematics. *Synthese*, 125(3):341–383, 2000.
M. Redmayne. Objective probability and the assessment of evidence. *Law, Probability and Risk*, 2:275–294, 2003.
R. Richards. *The Species Problem. A Philosophical Analysis*. Cambridge University Press, Cambridge, 2010.
F. S. Roberts. *Measurement Theory*. Addison-Wesley, Reading MA, 1979.

F.A. Rodriguez-Consuegra. *The Mathematical Philosophy of Bertand Russell: origins and development*. Birkhäuser, 1991.

C.S. Roero. *Peano e la sua scuola*. Deputazione Subalpina di Storia Patria, 2010.

H. Rogers. *Theory of recursive functions and effective computability*. MIT Press, Cambridge MA, 1987.

D. Rosen. Vicariant patterns and historical explanation in biogeography. *Systematic Zoology*, 27:159–188, 1978.

G. Rosen. The reality of mathematical objects. In J. Polkinghorne, editor, *Meaning in Mathematics*, pages 113–131. Oxford University Press, Oxford, 2011.

D. Rowbottom. Scientific progress without increasing verisimilitude: In response to Niiniluoto. *Studies in History and Philosophy of Science*, 51:100–104, 2015.

R. Rudner. Value judgments in scientific validation. *Scientific Monthly*, 79(3):151–153, 1954.

E. Runggaldier. Are there 'tensed' facts? In F. Stadler and M. Stltzner, editors, *Time and History. Proceedings of the 28th International Ludwig Wittgenstein Symposium*, pages 77–84. Ontos, 2006.

B. Russell. Sur la logique de relations avec des applications á la théorie des series,. *Rivista di Matematica*, VII:115–148, 1901.

B. Russell. *The Principles of Mathematics*. Cambridge University Press, 1903.

B. Russell. *Our Knowledge of the External World as a Field for Scientific Method in Philosophy,*. Open Court, 1914.

B. Russell. *History of Westem Philosophy*. George Allen and Unwin, London, 1946.

B. Russell. *Towards the "Principles of Mathematics"*, volume III of *Collected Papers of Bertrand Russell*. Routledge, 1993.

N. Salmon. The limits of human mathematics. *Nous*, 35:93–117, 2001.

W.C. Salmon. Partial entailment as a basis for inductive logic. In N. Rescher, editor, *Essays in Honor of Carl G. Hempel*, chapter 47–82. Springer, Dordrecht, 1969. Reprinted as "The 'partial entailment' theory in *Reality and Rationality*", Oxford, Oxford University Press, ch. 11, 184–209.

J. Schaffer. On what grounds what. In D. Manley, D.J. Chalmers, and R. Wasserman, editors, *Metametaphysics: New Essays on the Foundations of Ontology*, pages 347–383. Oxford University Press, Oxford, 2009.

A.F. Schmid, editor. *Bertrand Russell. Correspondance sur la philosophie, la logique et la politique avec Louis Couturat (1897–1913)*. Editions Kimé, 2001.

H. Scholz. Wie ist eine evangelische theologie als wissenschaft möglich? *Zwischen der Zeiten*, 9:8–53, 1931.

H. Scholz. Was ist unter einer theologischen aussage zu verstehen? In *Theologische Aufsätze. Karl Barth zum 50. Geburtstag*, pages 25–37. Kaiser, München, 1936.

H. Scholz and H. Schweitzer. *Die Sogennanten Definitionen durch Abstraktion*. Felix Meiner, 1935.

G. Schurz. Relevant deduction. In W. Spohn, editor, *Erkenntnis Orientated: A Centennial Volume for Rudolf Camap and Hans Reichenbach*, pages 391–437. Springer, Dordrecht, 1991.

G. Schurz. Probabilistic default reasoning based on relevance and irrelevance assumptions. In D. Gabbay, editor, *Qualitative and Quantitative Practical Reasoning*, pages 536–553. Springer, 1997.

G. Schurz. Probabilistic semantics for delgrande's conditional logic and a counterexa- mple to his default logic. *Artificial Intelligence*, 102(1):81–95, 1998.

G. Schurz. Verisimilitude and belief revision. with a focus on the relevant element account. *Erkenntnis*, 75:203–221, 2011.

G. Schurz and P. Weingartner. Verisimilitude defined by relevant consequence-element. In T.A.F. Kuipers, editor, *What is Closer-to-the-truth?*, pages 47–77. Rodopi, Dordrecht, 1987.

G. Schurz and P. Weingartner. Zwart and franssen's impossibility theorem holds for possible-world-accounts but not for consequence-accounts to verisimilitude. *Synthese*, 172:415–436, 2010.

P. Schweizer. A syntactical approach to modality. *Journal of Philosophical Logic*, 24:451–454, 2002.

D. Scott. Axiomatizing set theory. In L. Henkin, editor, *Proceedings of Symposia in Pure Mathematics*, volume XIII. AMS, Providence, 1974.

J. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3:417–457, 1980a.

J. Searle. Intrinsic intentionality. *Behavioral and Brain Sciences*, 3:450–456, 1980b.

J. Searle. *Intentionality. An essay in the philosophy of mind*. Cambridge University Press, Cambridge, 1983.

J. Searle. Yin and Tang strike out. In D.M. Rosenthal, editor, *The Nature of Mind*, pages 525–526. Oxford University Press, New York, 1991.

J. Searle. *The Rediscovery of the Mind*. MIT Press, Cambdrige (Mass), 1992.

J. Searle. *The Construction of Social Reality*. Penguin, London, 1995.

J. Searle. The chinese room. In R.A. Wilson and F. Keil, editors, *MIT Encyclopaedia of the Cognitive Sciences*, pages 115–116. MIT Press, Cambdrige (Mass), 1999.

J. Searle. Twenty one years in the chinese room. In J. Preston and M. Bishop, editors, *Views into the Chinese Room: Essays on Searle and Artificial Intelligence*, pages 51–59. Clarendon Press, Oxford, 2002.

J. Searle. What is to be done? *Topoi*, 25:101–108, 2006.

P. Setoh, D. Wu, R. Baillargeon, and R. Gelman. Young infants have biological expectations about animals. *Proceedings of the National Academy of Sciences*, 110(40):15937–15942, 2013.

S. Shapiro. Incompleteness, mechanism, and optimism. *Bulletin of Symbolic Logic*, 4:273–302, 1998.

S. Shapiro. Incompleteness and inconsistency. *Mind*, 111:817–832, 2002.

S. Shapiro. Mechanism, truth, and Penrose's new argument. *Journal of Philosophical Logic*, 32: 19–42, 2003.

S. Shelah. The future of set theory. In H. Judah, editor, *Set Theory of the Reals*, volume 6, pages 1–12. Israel Mathematical Conference, 1993.

S. Shelah. Logical dreams. *Bulletin of the American Mathematical Society*, 40(2):203–228, 2003.

S. Shoemaker. *Physical Realization*. Oxford University Press, Oxford, 2007.

G.G. Simpson. *Principles of Animal Taxonomy*. Columbia University Press, New York, 1961.

P. Sloan. Buffon, German biology, and the historical interpretation of biological species. *The British Journal for the History of Science*, 12(2):109–153, 1979.

E. Sober. Evolution, population thinking, and essentialism. *Philosophy of Science*, 47(3): 350–383, 1980.

E.S. Spelke, A. Phillips, and A. Woodward. Infants' knowledge of object motion and human action. In D. Sperber, D. Premack, and A.J. Premack, editors, *Causal Cognition. A Multidisciplinary Debate*, pages 44–78. Clarendon Press, 1995.

D. Stamos. Species, languages, and the horizontal / vertical distinction. *Biology and Philosophy*, 17:171–198, 2002.
D. Stamos. *The Species Problem*. Lexington Books, Oxford, 2003.
D. Stamos. *Darwin and the Nature of Species*. State University of New York Press, Albany, 2007.
J. Steel. Mathematics needs new axioms. *The Bulletin of Symbolic Logic*, 4:422–433, 2000.
J. Stern. Modality and axiomatic theories of truth II: Kripke-feferman. *Review of Symbolic Logic*, 7:299–318, 2014.
J. Stern. *Towards a predicate approach to modal notions*. Springer, Berlin-New York, 2015.
H. Steward. Animal agency. *Inquiry*, 52(3):217–231, 2009.
O. Stolz. *Vorlesungen über allgemeine Arithmetik*. Teubner, 1885.
G. Strawson. Realism and causation. *The Philosophical Quarterly*, 37(148):253–277, 1987.
P. Suppes. Probabilistic inference and the principle of total evidence. In J. Hintikka and P. Suppes, editors, *Aspects of Inductive Logic*, pages 49–65. North-Holland, Amsterdam, 1965.
L. Tambolo. A tale of three theories: Feyerabend and popper on progress and the aim of science. *Studies in History and Philosophy of Science*, 51:33–41, 2015.
A. Tarski. The semantic conception of truth and the foundations of semantics. In H. Feigl und W. Sellars, editor, *Readings in Philosophical Analysis*, pages 52–84. Appleton-Century Crofts, New York, 1949.
I. Thalberg. How does agent causally work? In M. Brand and D.N. Walton, editors, *Action Theory. Proceedings of the Winnipeg Conference on Human Action*, pages 213–238. Reidel, 1980.
P. Tichy. On Popper's definitions of verisimilitude. *Studies in History and Philosophy of Science*, 25(2):155–160, 1974.
R. Tieszen. *Mathematical Intuition*. Springer, Dordrecht, 1989.
R. Tieszen. Gödel's path from incompleteness theorem (1931) to phenomenology (1961). *Bulletin of Symbolic Logic*, 4:181–203, 1998.
E. Topitsch. *Vom Ursprung und Ende der Metaphysik. Eine Studie zur Weltanschauungskritik*. Springer, Wien, 1958.
E. Topitsch. *Heil und Zeit. Ein Kapitel zur Weltanschauungsanalyse*. Mohr (Siebeck), Tübingen, 1990.
E. Troeltsch. Die absolutheit des christentums und die religionsgeschichte (1902, second edition 1912). In *Kritische Gesamtausgabe*, volume 5. de Gruyter, Berlin-New York, 1998.
A.M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.
G. Vailati. Dipendenza fra le proprietà delle relazioni. *Rivista di Matematica*, II(161–164), 1892.
J. Van Benthem and F. Velazquez-Quesada. The dynamics of awareness. *Synthese*, 177:5–27, 2010.
B. Van Fraassen. Singular terms, truth-value gaps and free logic. *Journal of Philosophy*, 63: 481–495, 1966.
J.D. Velleman, editor. *The Possibility of Practical Reason*. Oxford University Press, 2000.
P. Vrana and W. Wheeler. Individual organisms as terminal entities: Laying the species problem to rest. *Cladistics*, 8:67–72, 1992.
Vuillemin. *La Logique et le Monde Sensible. Étude sur les théories contemporaines de l'abstraction*. Flammarion, 1971.
H. Wang. *A Logical Journey*. MIT Press, Cambdrige (Mass), 1996.
H. Weber. Elementare mengenlehre. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, XV:173–184, 1906.
M. Weber. *Wissenschaft als Beruf*. Duncker & Humblot, München-Leipzig, 1919.

Z. Weber. A paraconsistent model of vagueness. *Mind*, 119:1026–1045, 2010.

C. Weidemann. *Die Unverzichtbarkeit natuürlicher Theologie*. Number 129 in Symposion. Alber, Freiburg, 2007.

E. Weydert. Rational default quantifier logic. In D. et al. Gabbay, editor, *Qualitative and Quantitative Practical Reasoning*, LNAI 1244, pages 589–599. Springer, Berlin, 1997.

H. Weyl. Über die definitionen der mathematischen grundbegriffe. *Mathematisch-naturwissenschaftliche Blätter*, 7:93–95; 109–113, 1910.

E.O. Wiley. *Phylogenetics*. John Wiley & Sons, Inc., New York, 1981.

J. Wilkins. *Species. The History of the Idea*. University of California Press, Berkeley, 2009.

T. Winograd. Understanding, orientations, and objectivity. In J. Preston and M. Bishop, editors, *Views into the Chinese Room: Essays on Searle and Artificial Intelligence*, pages 80–94. Clarendon Press, 2002.

L. Wittgenstein. *Philosophical Investigations*. Macmillan, New York, 1956. Edited by G. E. M. Anscombe.

L. Wittgenstein. *Remarks on the Philosophy of Psychology*. University of Chicago Press, Chicago, 1958. Edited by G.E.M. Anscombe and G.H. von Wright.

W.H. Woodin. The continuum hypothesis I. *Notices of the American Mathematical Society*, 48: 567–576, 2001.

S. Yablo. Mental causation. *Philosophical Review*, 101:245–280, 1992.

S. Zwart. *Refined Verisimilitude*. Kluwer, Dordrecht, 2001.

S. Zwart and M. Franssen. An impossibility theorem for verisimilitude. *Synthese*, 158(1):75–92, 2007.