Truth and Proof in the Long Run

ACTA PHILOSOPHICA GOTHOBURGENSIA 29

Truth and Proof in the Long Run

Essays on Trial-and-Error Logics

Martin Kaså

UNIVERSITY OF GOTHENBURG
ACTA UNIVERSITATIS GOTHOBURGENSIS

ISBN 978-91-7346-903-6 (PRINT)
ISBN 978-91-7346-904-4 (DIGITAL)
ISSN 0283-2380

Available online at: http://hdl.handle.net/2077/51792

Distribution:
ACTA UNIVERSITATIS GOTHOBURGENSIS
Box 222, 405 30 Göteborg, Sweden
acta@ub.gu.se

Typeset in Adobe Garamond Pro using XeLaTeX
Source edited in GNU Emacs on Ubuntu GNOME GNU/Linux
Cover design by Peter Johnsen, using the MOONSHINER font by Mattox Shuler

Printed by Ineko, Kållered 2017

Abstract

Title: Truth and Proof in the Long Run. Essays on Trial-and-Error Logics
Author: Martin Kaså
Doctoral thesis in theoretical philosophy
Language: English
ISBN: 978-91-7346-903-6 (Print)
ISBN: 978-91-7346-904-4 (Digital)
ISSN: 0283-2380
Keywords: convergence, dynamic meaning, experimental logics, knowable consistency, tableaux systems, trial-and-error

The theme of this book is convergence. For many philosophical representations of the evolution of theories, as well as representations of the meaning of the language used to express these theories, it has been essential that there exists some kind of convergence. This thesis introduces and collects four papers in philosophical logic pertaining to two different aspects of this basic tenet. On one hand, we have theories, their axioms and their rules of inference. We often have reason to revise a theory over time, to delete some axioms, add some new ones, or perhaps even revise our modes of reasoning. A simple model of such activity, providing a definition of what it may mean that something is *provable in the long run* in such a dynamic setting, is here investigated, and its relevance for the philosophical discussion about mechanism and knowable self-consistency is evaluated. On the other hand, the notion of a convergent concept, a term which, for whatever reason, has a certain tendency to its application over time, gets a precise explication in terms of *trial-and-error classifiers*. Formal languages, based on these classifiers, are introduced with semantics and proof systems, and are explored using standard logical methods.

This doctoral thesis is based on the following papers.

I Experimental Logics, Mechanism and Knowable Consistency
Martin Kaså
Originally published in *Theoria*, 78(3):213–224, 2012
doi: 10.1111/j.1755-2567.2012.01133.x

II A Logic for Trial and Error Classifiers
Martin Kaså
Originally published in *Journal of Logic, Language and Information*, 24(3):307–322, 2015
doi:10.1007/s10849-015-9222-7

III Formally Modelling Convergent Dynamic Meaning. Results on Compactness and Axiomatizability
Martin Kaså
Submitted 2016

IV Analytic Tableaux for Trial-and-Error Reasoning
Martin Kaså
Manuscript 2017

Previously published papers are reprinted with permission.

Acknowledgements

First, not being mentioned here does not mean I don't owe you! There are many more colleagues who collectively and individually make my working life rather agreeable. Those referred to by name here have been identified as having had the most obvious, be it direct or indirect, effect on my thesis work. Mostly a positive effect, but perhaps not invariably so.

To my thesis advisors: Christian Bennet, you are the *sine qua non* of this project, from its very inception. Fredrik Engström, you have been with me as both a colleague and a friend through every technical, and almost every psychological, step of this work. Dag Westerståhl, I just can't imagine I would ever have finished this book without you. Heartfelt thanks!

To my former colleagues in Lund: Nils-Eric Sahlin, you started me off once, and your passionate introduction to Ramsey's philosophy left a profound mark on all my philosophical thought. Johannes Persson, you were one of my first teachers, I have always admired your keen philosophical common sense, and I still have hopes for us writing something together. Mats Johansson, my co-author of totally unrelated philosophical work; those were the days! Linus Broström, you were so friendly and helpful during my previous, not-so-productive, time as a PhD student working in other areas of philosophy.

To colleagues from the old Philosophy department: Alexander Almér, you are very much missed since you moved to the IT Faculty, but I will have ample free time now, so maybe we can have some real collaboration between our departments. Björn Haglund, you are, from where I am standing, the grand old man of Gothenburgian philosophy, and I hope you will enjoy seeing this work in print. Martin Filin Karlsson, please just come back and do more philosophy already!

To my colleagues and friends at CLT and Språkbanken: Thank you for accepting me as your adopted son in the language technology community for a few years. It was fun, I learned a lot, and, as you know, when I miss

you too much I just come back. In particular I want to mention Yvonne Adesam, Lars Borin, Markus Forsberg and Nina Tahmasebi.

Among international colleagues, Dora Achourioti deserves special mention for inviting me to Amsterdam, at a crucial point in time, to discuss common research interests.

To my colleagues at FLoV: Peter Johnsen, thank you for the lovely cover design for this book, and for very much more besides. Felix Larsson, we teach, we talk, you really make my position as a lecturer seem like a good choice of lifestyle. Stellan Petersson, our discussion about Putnam's ideas on meaning was invaluable, and the raised fist salute in the corridors provides me with much needed energy. Anna-Sofia Maurin, you believed me capable of finishing, for some reason. Rasmus Blanck, you are so helpful, in so many ways, that you have almost made it fun wrapping up this thesis work. Anton Broberg, thanks for letting me bounce some half-baked ideas off your sharp mind. Thanks to the administrative staff, and in particular to Tobias Pettersson and Madelaine Miller, for making it possible to work as a teacher at our department. I would also like to thank, quite generally, all colleagues who have taken part in the seminars in theoretical philosophy and logic, where I have occasionally presented drafts of conference papers, journal articles, and this thesis.

To my friends and family: I love you all, but here I am just going to mention those who have suffered most from my tardiness in finishing this work: Sophia, Viggo, Hannes. I am at a loss for words.

Funding from the foundation *Kungliga och Hvitfeldtska Stiftelsen* is hereby gratefully acknowledged.

Martin Kaså, Göteborg, February 2017

Contents

No this is how it works
You peer inside yourself
You take the things you like
Then try to love the things you took

Regina Spektor

1 Introduction

Wherein we get some general guidelines on how to approach this thesis; what it is, what it is not, how it is supposed to be read.

1.1 Methodology

As sentient beings, we are presented with a world of objects—a vast and complicated abundance of them—and try collectively to make some sense of our surroundings. We do this partly by putting and (temporarily) storing the things in virtual conceptual boxes or categories, i.e., we classify. I am convinced that valuable classification, meaning classification that is useful for systematization, prediction and explanation, is possible due to a complex interplay of many different aspects: of regularities in the world, our conceptual powers, the sophistication of our instruments, and also such things as our communicative abilities and intersubjective conceptual schemes. The metaphysical structure of the world does not on its own constitute a sufficient ground for this possibility.[1]

I arguably belong to some kind of instrumentalist and pragmatist philosophical camp, and I confess to a deeply felt skepticism towards the notion that reality presents us with *natural kinds*, to which we can rigidly point using terms in our (scientific) language.[2] This skepticism notwithstanding, it is hard to deny that some kind of *convergence* over time in the extension of the terms of our language (and our body of beliefs) is of utmost importance, when it comes to scientific theorizing, and also for language as a communicative device in general. From my position, it is obviously not a viable option to just say that "gold" rigidly refers to the substance *gold*, and

[1] See (Kaså, 2015).

[2] Cognoscenti will recognize this statement as opposed to the main point of (Putnam, 1975)—a very well-written and inspiring work, which will be briefly reviewed in Section 2.2.2, and returned to in Section 3.2.1.

therefore even has a *constant* extension over time whether we realize it or not. My preferred understanding of convergence must be, in some sense, pragmatically determined.[3]

The position I am starting from could perhaps be quickly summed up by saying: (i) the extension of a term may vary over time, (ii) whether there will be convergence, whether the term will turn out to be useful for predictions, inductive generalizations, etc., is not, or at least not solely, determined by metaphysical features of the world, but may be largely "accidental" and dependent on a whole interconnected web of other terms (scientific and otherwise), which are also evolving, and (iii) even if there is convergence, there may never be an actual point in time when a given term reaches a final, unchanging extension.

Now, obviously, a lot more will be said about this, and at another level of precision, in Chapters 2 and 3, but there it is; this is the basic idea, my philosophical point of departure. This is a book about convergence.

Where to go from there? Let me be upfront with what I have *not* done in this thesis. I have not done much classical "conceptual analysis", not really analytically explored this shard of a philosophical position. Neither have I engaged in earnest with the vast literature on natural kinds (in and out of philosophy of science), nor with the more linguistically oriented literature about lexical change. Moreover, I have not tried to directly apply my particular convergence concept to philosophical problems concerning the use of scientific terms over time. All these things are certainly interesting and worthwhile, but, instead, the attitude of this thesis is, I guess, fairly typical for a philosophical logician working in what could be called an "exploratory" mode. Cautiously and meticulously, I have just wanted to know exactly what I am talking about, to make the concepts involved as formally precise as I can, and in particular investigate what inferential properties these concepts have.

[3]On an autobiographical note, there can be little doubt that the origins of my thoughts on many of the philosophical issues touched upon in this book can be causally traced to the presentation of F. P. Ramsey's pragmatism in (Sahlin, 1990). In *Facts and propositions*—a paper which is still today a delight to read—Ramsey says that "The essence of pragmatism I take to be this, that the meaning of a sentence is to be defined by reference to the actions to which asserting it would lead, or, more vaguely still, by its possible causes and effects." (Ramsey, 1927, p. 57)

Starting from a rough idea about convergence of a rather weak variety (to be presented in Section 3.2.1) of terms' extensions over time, my methodology is, I think, heavily inspired by Carnap's notion of explication.

> The task of *explication* consists in transforming a given more or less inexact concept into an exact one or, rather, in replacing the first by the second. We call the given concept (or the term used for it) the *explicandum*, and the exact concept proposed to take the place of the first (or the term proposed for it) the *explicatum*. The explicandum may belong to everyday language or to a previous stage in the development of scientific language. The explicatum must be given by explicit rules for its use, for example, by a definition which incorporates it into a well-constructed system of scientific either logicomathematical or empirical concepts. (Carnap, 1950, p. 3)

So the, perhaps half-baked, idea about "convergent concepts" from my informal semantics is in my technical work replaced by what will be called "trial-and-error classifiers" in this thesis. The "well constructed system" is simply a formal language with syntax and semantics given in standard, and rather elementary, logical terms. Now, as soon as we have a formal semantics, and (thereby) a *logic*, there are a host of natural questions which need to be addressed.[4] Is reasoning in the logic mechanizable? Is it always finitary? Is there an algorithm for finding interpretations of satisfiable formulas? How does this logic compare to standard logics? Are there interesting fragments or extensions? Pursuing this kind of questions mostly takes the shape of open-minded investigation. The result of the explication, what Carnap calls the explicatum, is an exactly defined concept, and we want to know more about it. As it happens, the concept is logical, and hence there is a very useful set of tools to equip ourselves with for the exploratory journey.[5]

[4] "Natural" and "need to" are in this case basically instinctive logician's judgements. But the general sentiment extends to all branches of philosophy. A mere definition is not enough to really get to *know* a concept; we also want to investigate the consequences (and presuppositions) of the chosen definition.

[5] Admittedly, the tools have to be non-trivially adjusted to the particular problem at hand.

If trial-and-error classifiers are thought of as the result of a process of explication, is the explicatum any good? Carnap (1950, p. 7) says:

> 1. The explicatum is to be *similar to the explicandum* in such a way that, in most cases in which the explicandum has so far been used, the explicatum can be used; however, close similarity is not required, and considerable differences are permitted.
> 2. The characterization of the explicatum, that is, the rules of its use (for instance, in the form of a definition), is to be given in an *exact* form, so as to introduce the explication into a well-connected system of scientific concepts.
> 3. The explicatum is to be a *fruitful* concept, that is, useful for the formulation of many universal statements (empirical laws in the case of a nonlogical concept, logical theorems in the case of a logical concept).
> 4. The explicatum should be as *simple* as possible; this means as simple as the more important requirements (1), (2), and (3) permit.

I submit that it remains to be seen whether (3) is true in the present case. The situation seems promising, but without doubt, much honest philosophical toil will be required to get, even nearly, conclusive evidence for this.[6]

The way of applying logical methods to problems in philosophy described above can, somewhat vaguely, be characterized as being *positive* in spirit; a creation of new formal systems to represent informal, philosophically interesting, concepts. Not all instances of philosophical logic are of this variety, and the present thesis also exemplifies another brand. From the very outset the two papers (Putnam, 1965) and (Jeroslow, 1975), though they ostensibly address quite different problems than the problem of dynamic meaning (*viz.*, generalization into the trial-and-error dimension of

[6]And there may be the worry that, though I am certainly otherwise convinced, we have a case here where the explicatum will eventually fail to be fruitful, since the explicandum itself is misguided.

computability and provability, respectively), have constituted perhaps the most important inspiration for my work. They are also very interesting in their own right and, along the way, I stumbled over an attempt to use (Jeroslow, 1975) in devising a counter-argument against a certain type of arguments aimed at demonstrating the (mathematical) impossibility of representing mind as a machine.[7] I found the suggestion in several ways interesting, sharpened it a bit with some technical work, but in the end found the line of reasoning unconvincing.[8]

While this latter work of mine could be considered a detour with respect to the main investigation, it was, I think, fruitful to engage with the concept of convergence from a different angle; to directly address theories rather than languages. And while largely "negative", there is also here a strong sense of exploration. Jeroslow's systems have not, in my opinion, been sufficiently investigated by philosophically minded logicians. There are plenty more possible questions in connection with these concepts which I, for one, would like to see both formulated and answered.[9]

1.2 Layout

After the short general introduction in this chapter, the body text of the thesis is structured in two separate main parts. First, there is the background Chapter 2, aimed at giving the research presented in the four papers (Kaså, 2012, 2015, 2016, 2017) some proper context. Typically, I have tried to give the issues a somewhat fuller (and wordier) presentation than is possible within the stylistic and spatial confines of a journal article. And this is not just to better set the philosophical stage, but also to give some technical material from mostly (Jeroslow, 1975), necessary for a proper understanding of some of my work, but perhaps not all that well known.

[7] The standard reference for the start of this debate is the oft-cited (and overwhelmingly critically so) (Lucas, 1961). The above-mentioned attempt to use Jeroslow which I first came across was in (Hazen, 2006a), though others seem to have been on a similar track.

[8] See Section 3.1 and (Kaså, 2012). This is not to say that I accept Lucas's anti-mechanist arguments. On the contrary, I am convinced that they are erroneous, but this is based on other considerations than the attempt to use (Jeroslow, 1975).

[9] Some pointers to recent literature are given in Section 2.1.3.

The background is further thematically divided into two sections, covering two different, but interconnected, aspects of convergence. On the one hand, we look at convergence in *theories*. As preparatory work for reading Section 3.1 and my paper (Kaså, 2012), the most important thing in this Section 2.1 is the technical description of the so-called *experimental logics* of (Jeroslow, 1975), which are simple models for dynamic axiomatic theories, where the axioms, or rules of inference, may change over time. These systems give a meaning, albeit a rather simplistic one, to the concept "provable in the long run", somewhat like the semantics alluded to below gives a meaning to "true in the long run". In my thinking, these experimental logics have become inextricably linked with discussions about some philosophical arguments about the (im)possibility of a true mechanist account of (some of the faculties of) the human mind, and different ways of interpreting what it means for a set to be "computably produced". And, as it turns out, more people think this way, so there will also be some background from sources such as (Boolos, 1995), (Lucas, 1961), (McCarthy and Shapiro, 1987), (Shapiro, 1998), and others.

On the other hand, we have our focus on the phenomenon (or, rather, phenomena) of convergence in the languages that theories are couched in. That is, we investigate how terms may function over time and perhaps mean (or denote) different things at different times, but have a potential to stabilize in meaning. Issues like this are perhaps of special significance when it comes to scientific theorizing, and the most important reference here is (Putnam, 1975), but there will also be some background from (Peirce, 1877) and (Peirce, 1878). Thus Section 2.2 will set the stage for a reading of my work on *trial-and-error logics*, presented in Section 3.2, and originally in the three papers (Kaså, 2015, 2016, 2017). One may say that this second theme is, by and large, *semantic* in nature.

After the background and context is given, there is the aptly named Chapter 3: Contributions. Herein the actual new results and discussions from my papers are presented thematically. Full proofs of technical results are, in general, not given in this chapter; for that, the reader is referred to the original papers.

First, in Section 3.1, it is argued that, while interesting and valuable in many ways, Jeroslow's experimental logics cannot in any definitive way be used as the final word in the discussion on (anti-)mechanism. Which is not to say that thinking along lines like these cannot shed new light on the debate. Some results from Jeroslow are made clearer, and also extended.[10]

Then, there is Section 3.2, where my own formal take on convergence in dynamic meaning is given a technical presentation. This includes a brand new semantics for a syntactically familiar language, and proofs of properties such as axiomatizability and compactness. A natural fragment of this language is also distinguished, and for this we give two proof systems, natural deduction and analytic tableaux, respectively, which are proven to be sound and complete. Along the way we get the (expected) result that the fragment is decidable, while the full language is not.

The main parts are followed by the short Chapter 4, basically listing some open problems, mostly of a technical character, which seem to point towards reasonable, and hopefully fruitful, directions of further research in this area. While this could alternatively have been included in Chapter 3, this mode of presentation has the added value of making the problems more salient, and everyone is cordially invited to partake in the quest for definitive solutions, as well as the formulation of even harder questions.

For the convenience of casual readers, the thesis ends with brief summaries of the four research papers. Though it is of course preferred that the papers themselves be read—since that is where the research contributions really take place—this Chapter 5 at least gives an indication of how the research thematically presented in Chapter 3 in fact has been, and will be, published. What the reader will find there are essentially extended abstracts.

The original papers are, as customary, attached to the introductory text of the thesis, in published or manuscript form.

[10]As will become clear in this chapter, and in the paper (Kaså, 2012) itself, I owe a lot to Allen Hazen for even starting to think about these matters. See (Hazen, 2006a,b).

2 Logico-philosophical strands

The present chapter gives an historical background (or rather backgrounds in the plural) to the logical investigations in the papers of this thesis. When presenting a compilation of papers, one can always worry about the degree of cohesion; what is it that makes this *one* project rather than several disparate ones? But it should be clear by now that there really is *one* overarching theme here: convergence. This will be reinforced by the following review.

2.1 Semi-Euclidean theories, Δ^0_2 and consistency

This section is devoted to introducing the mainly technical background for the part of the thesis which is about convergence in *theories*, and the connection between a formal representation of this phenomenon and some philosophical questions regarding mechanical models of thinking.

2.1.1 Jeroslow's experimental logics

The mission statement of Jeroslow's *Experimental logics and Δ^0_2-theories* is this:

> In this paper, we explore the concept of a logic which proceeds by trial-and-error, and deduce consequences which follow from relatively weak assumptions about these *experimental logics*.
> (Jeroslow, 1975, p. 253)

Just reading this, one could perhaps expect something like the semantically motivated trial-and-error logics of the present thesis (Section 3.2), but though Jeroslow's work may definitely be related in spirit (and pedigree) to such considerations, it is in fact very different. I start this section with a

résumé of the technicalities of Jeroslow's article, which underlie the work presented in Section 3.1.[11]

When Jeroslow talks about an "experimental logic" he technically just means a recursive, or decidable, ternary relation $H(t, p, \varphi)$ on the set of natural numbers. What motivates the terminology is the intended interpretation, which is:

"At time t, the construction p is recognized as a proof of φ."[12]

How is this a "logic which proceeds by trial-and-error"? The answer lies in the definition of what it means to be provable in an experimental logic. The set of theorems of H, denoted by $\mathrm{Th}(H)$, is taken to be the set Rec_H of *recurring formulas*, defined by:

$$\mathrm{Rec}_H(\varphi) \Leftrightarrow \forall s \exists t > s \exists p H(t, p, \varphi)$$

So we get a picture of a dynamic, or *evolving* theory: whether it is because we change axioms or rules of inference, and whatever reasons we have for doing so, different formulas may be provable at different points in time. And the "real" theorems are the recurrent ones, the ones we never permanently throw away. Complexity-wise, this is a Π^0_2-set. Remember that H is recursive, so this model would not fit a situation where we, e.g., arrive at axioms from non-computable sources, as divine inspiration and the like. To quote the originator again:

> The experimental logics we study here are the most conservative extension of formal systems into the trial-and-error dimension, since we hypothesize that the events which may cause changes in axioms and rules of reasoning are mechanical, and the reformulation of the theory following these events is also mechanically determined. (Jeroslow, 1975, p. 254)

[11] Meta-mathematical terminology and tools used here are standard, and the reader is referred to, e.g., (Lindström, 1997) for details.

[12] Here, and in other applicable cases, no distinction is made between syntactic objects, their Gödel numbers, and the corresponding numerals.

Jeroslow is not particularly interested in Π^0_2-sets in general, but limits himself to what he calls *convergent* experimental logics. Some formulas may be not only infinitely recurring, but in fact stabilize as being always provable from some point in time. The Σ^0_2-set Stbl_H of *stable* formulas is defined by

$$\mathrm{Stbl}_H(\varphi) \Leftrightarrow \exists p \exists s \forall t > s H(t, p, \varphi)$$

An experimental logic is defined to be convergent if, for all φ, we have that $\mathrm{Rec}_H(\varphi) \rightarrow \mathrm{Stbl}_H(\varphi)$, if every theorem is "decided in the limit", as it were. So, by definition, the set of theorems of a convergent experimental logic is actually a Δ^0_2-set.

In his paper (1965), Putnam defined X to be a one-place *trial and error predicate* if there exists a recursive function f such that for all $n \in \omega$:

$$\begin{cases} n \in X & \Leftrightarrow & \lim_{m \to \infty} f(n, m) = 1 \\ n \notin X & \Leftrightarrow & \lim_{m \to \infty} f(n, m) = 0 \end{cases}$$

His first characterization theorem then states that X is a trial and error predicate iff $X \in \Delta^0_2$. (Putnam, 1965, p. 51) So given any such X (and f), we can define H by e.g., $H(t, p, \varphi) \Leftrightarrow (f(\varphi, t) = 1 \wedge p = 0)$. This evidently makes H a convergent experimental logic, and $\mathrm{Th}(H) = X$. From this observation we get the following characterization:

Theorem 1 (Jeroslow/Putnam)**.** *The sets of theorems of convergent experimental logics are precisely the* Δ^0_2*-sets.*

Next, extending Gödel's first incompleteness theorem, Jeroslow provides a short proof of the basic incompleteness theorem for experimental logics:

Theorem 2 (Jeroslow)**.** *If* H *is a consistent, convergent, experimental logic which contains first-order Peano arithmetic, and is closed under deduction, then* $\mathrm{Th}(H)$ *is incomplete, even at the* Π^0_1*-level.*

Proof. From the fixed-point lemma, we know that there is a sentence φ such that $\mathrm{PA} \vdash \varphi \leftrightarrow \neg \mathrm{Stbl}_H(\varphi)$. Consider the two cases:

1. $\varphi \in \mathrm{Th}(H)$. Since H is convergent, $\mathrm{Stbl}_H(\varphi)$ is true, but can, given the fixed point, not be a theorem of H on pain of contradiction.

2. $\varphi \notin \mathrm{Th}(H)$. Use the fact that $\vdash \mathrm{Stbl}_H(\varphi) \rightarrow \mathrm{Rec}_H(\varphi)$. Then $\neg\mathrm{Rec}_H(\varphi)$, which is true, cannot be a theorem, since $\neg\mathrm{Stbl}_H(\varphi)$ would be, and hence also φ, contradicting the case assumption.

In any case, there would be a true, "unprovable" formula equivalent to a Σ^0_2-sentence $\exists x\psi(x)$. But then there exists a number n for which $\psi(n)$ is a true, unprovable Π^0_1-sentence.[13] □

The upshot is that even though the concept of a *theorem* is more complex for experimental logics than for ordinary formal theories (Δ^0_2 rather than Σ^0_1) the incompleteness phenomenon still occurs at the lowest possible level, viz., Π^0_1, so there are still "real" (in Hilbert's sense) true mathematical propositions which cannot be reached even through such an infinite, mechanistic, trial-and-error process which can be represented as an experimental logic.[14] Note, though, that we have not explicitly given an actual Π^0_1-sentence, and this is not by accident.

When it comes to Gödel's second theorem, the incompleteness phenomenon is relativized. For an ordinary formal theory, such as first-order Peano arithmetic, it is easy to mechanically find a true, unprovable Π^0_1-sentence; just take $\mathrm{Con}_{\mathrm{PA}}$. In contrast, observe the following theorem.

Theorem 3 (Feferman/Jeroslow)**.** *Some experimental logics prove their own consistency.*

Proof. Let T be an adequate, sound arithmetical theory (e.g., first-order PA) and let $\mathrm{Prf}_T(x, y)$ be the usual proof predicate representing (in T) that y is a proof of x.[15] Furthermore, let Con_T be the canonical consistency statement $\neg\exists y\mathrm{Prf}_T(\bot, y)$.

Define the experimental logic H as the following decidable predicate:

$$H(t, p, \varphi) \Leftrightarrow \begin{cases} \mathrm{Prf}_{T+\mathrm{Con}_T}(\varphi, p) & \text{and } \forall x \leq t \neg\mathrm{Prf}_{T+\mathrm{Con}_T}(\bot, x)) \text{ ; or} \\ \mathrm{Prf}_T(\varphi, p) & \text{and } \exists x \leq t \mathrm{Prf}_{T+\mathrm{Con}_T}(\bot, x)) \end{cases}$$

[13] This particular version of the proof is from (Bennet, 1989).

[14] The distinction between *real* and *ideal* mathematical propositions is spelled out in (Hilbert, 1925).

[15] It suffices that $\mathrm{Prf}_T(x, y)$ is *standard* in the sense of (Feferman, 1960).

Note that the set $T + \mathrm{Con}_T$ *is* consistent, so in fact we have that, for all t, $\exists p H(t, p, \varphi) \Leftrightarrow T + \mathrm{Con}_T \vdash \varphi$.

The following argument takes place inside $T + \mathrm{Con}_T$:

a) If $T + \mathrm{Con}_T \vdash \bot$, then $\mathrm{Th}(H) = \mathrm{Th}(T)$. But we have Con_T, and hence H is consistent.

b) If, on the other hand, $T + \mathrm{Con}_T \nvdash \bot$, then $\mathrm{Th}(H) = \mathrm{Th}(T + \mathrm{Con}_T)$, and, by the assumption, H is consistent.

This shows that $T + \mathrm{Con}_T \vdash$ "H is consistent", and we can conclude that the consistency of H is a theorem of H itself. □

This is my rendering of Jeroslow's proof, which in turn is just an adaption of Feferman's proof of Theorem 5.9 in (Feferman, 1960), from which Jeroslow states that he "abstracted the concept of an experimental logic".[16] Note two features of this which will become important in the later discussion:

- The canonical consistency statement of an experimental logic is not in general equivalent to a Π^0_1-sentence. If $\eta(t, p, \varphi)$ is a formula which serves as definition of the arithmetical relation H, then, using the definition of $\mathrm{Rec}_\eta(\varphi)$, we get a canonical Σ^0_2 consistency statement $\mathrm{Con}(\eta)$ defined by $\neg\forall s \exists t > s \exists p\, \eta(t, p, \bot)$.[17]

- The set $\mathrm{Th}(H)$ of the proof is actually Σ^0_1.

In fact, Jeroslow proved that we cannot in general effectively construct a true Π^0_1-sentence π such that $\pi \notin \mathrm{Th}(H)$, though we know that they exist. This is the content of Theorem 5 of his paper (the exact statement and proof is omitted here, since it is not essential for what is to come).

Finally, there is still a kind of "second incompleteness theorem", indicating that a class of experimental logics (satisfying some, rather reasonable, extra assumptions) cannot prove their own 2-consistency, and hence cannot prove their own soundness for certain trial-and-error statements.[18]

[16] Feferman used a (non-standard) "provability predicate", intensionally expressing provability in the largest consistent sub-theory of the original theory.

[17] Or, in case we assume convergence: the Π^0_2-formula $\forall t \forall p \exists s > t \neg\eta(s, p, \bot)$.

[18] (Jeroslow, 1975, p. 264)

Theorem 4 (Jeroslow)**.** *Suppose* $\mathrm{Th}(H) \supseteq \mathrm{PA}$ *is closed under deduction and* 1*-consistent, and furthermore:*

- *If* φ *is equivalent to a* Σ^0_2*-formula in* PA *then* $\vdash \mathrm{Rec}_H(\varphi) \to \mathrm{Stbl}_H(\varphi)$.
- *If* $\mathrm{PA} \vdash \alpha \to \beta$*, then* $\vdash \mathrm{Stbl}_H(\alpha) \to \mathrm{Stbl}_H(\beta)$.
- *If* $\rho \in \Sigma^0_1$*, then* $\vdash \rho \to \mathrm{Stbl}_H(\rho)$

Then H *cannot prove that it is* 2*-consistent.*

The proof can be found in (Jeroslow, 1975, pp. 264f).

2.1.2 Semi-Euclidean theories

Though the connection to Jeroslow has not always been explicitly noted, others have evidently been entertaining similar thoughts about generalizing the concept of a formal system. While assessing the alleged relevance of "limitative theorems", such as versions of Gödel's incompleteness theorems, for the philosophical debate on whether mind is (or can be, or can be represented as being) *mechanical*, Stewart Shapiro has this to say on a kind of idealization according to which the "product" of mathematics is just like a set of theorems of a formal system:

> The normative idealization is consonant with a longstanding epistemology for mathematics. The idea is that for mathematics at least, real humans are capable of proceeding, and should proceed, by applying infallible methods. In practice (or performance) we invariably fall short of this, due to slips of the pen and faulty memory, but in some sense we are capable of error-free mathematics. We start with self-evident axioms and proceed by gap free deduction. Call this the *Euclidean* model of mathematics. (Shapiro, 1998, p. 293)

This is of course but one of the possible "normative idealizations". What *are* we modelling, anyway? What kind of enterprise is it that should be represented (in some way) by some variety of formal system? The activity should be recognizable as human mathematical activity (albeit idealized),

and as such it must be fallible. And not only now, but forever fallible; it seems too much of a stretch to postulate that there will ever be an actual point in time where inconsistencies and other errors are just gone forever. It seems less far-fetched to idealize and say that *each individual error* will be spotted and corrected over time. Here is a picturesque description of a momentarily fallible, but dynamic and (in a sense) "eventually infallible" mathematical researcher:

Consider an ideal mathematician engaged in developing an axiomatic theory. She may perhaps start with some base theory which she does not question, and then she wants to extend it by adding new concepts and new pieces of information. Adding new concepts, in a formal setting, amounts to introducing new symbols to the language and adding axioms which characterize (as precisely as possible within the logical confines of the language) the intuitive mathematical ideas. Then, there is the business of deduction. Theorems pile up, and unwanted consequences may surface: contradictions, as well as formal theorems which, while perhaps consistent, show that the axiomatization fails to capture the intended informal concept. Luckily, our careful mathematician has kept track of which axioms were used in which proofs, and therefore she is in a position to backtrack, delete whatever axioms she holds responsible (as well as dependent theorems) and start anew. And thus the next step in the theory development is taken.

In a more fortunate case, she may be happy with the results so far, but she still wants to go on developing her theory by adding new axioms (concerning new or old non-logical symbols). Ideally, this process continues without limitations of space, time, etc. Not caring too much about what happens at each step, nor about why and how new axioms are chosen, we get a simple but reasonable picture of a discrete theory development over time. And the theory as a whole, the *dynamic theory*, is this entire sequence.

Something like this must be what Shapiro has in mind when he writes:

> [O]nce we leave the Euclidean model, even the ideal agents change their minds from time to time and so the model of a Turing machine printing out truth after truth is not appropriate. […] Suppose that whenever a human asserts a contradiction, or some other arithmetic falsehood, she has the

> ability in principle to realize the error and withdraw it. This assumption is a minimal retreat [...] Call it the *semi-Euclidean* assumption. (Shapiro, 1998, pp. 296f)

Let us call the set of sentences which are "eventually accepted" by such an idealized mathematician a *semi-Euclidean set*. There is nothing in the model in and of itself which precludes this set from being computably enumerable, and thus equivalent to the set of theorems of a formal system. But, and importantly, there seems to be no strong prima facie reason to believe that it should be. So a formal, mechanistic, model of this kind of mathematical activity has to cater for the possibility that the "product" of the system may be a set of (potentially) higher arithmetical complexity than the Σ^0_1 of a proper formal theory.

My presentation (and Shapiro's) has been in terms of mathematical theories, in order to provide an interface to Jeroslow's work and the later discussion in Section 3.1. But a similar story could arguably be told about other human scientific endeavours, and one could inquire into suitable representations of these. This is not far from Peirce's cautious optimism regarding convergence in the sciences, cf. the discussions in Section 2.2.1.

2.1.3 Related concepts and studies

In the preceding presentation, the studies by Putnam and Jeroslow were used as a point of departure. While not arbitrary, it should be admitted that this is at least to some degree coincidental. Other logicians have, sometimes independently, been working with very similar concepts, and the point of this short section is to summarily present at least a small sample from the literature, both contemporary with Putnam and Jeroslow, respectively, and a few more recent papers. Though this will not play any important part in what follows, it is included in the hope that it will be found useful to a reader who considers this area intriguing, and may want pointers to where to continue.

First off, published back-to-back in the very same volume with (Putnam, 1965) is an article by E. M. Gold, where he introduces essentially the same concept, under the label *limiting recursive* sets.

> A set, S, is called recursive if the questions "is $x \in S$" are decidable. S will be defined to be *limiting recursive* if these questions are decidable in the limit, i.e., if there is a guessing function $g(x, n)$, which is total recursive and such that, for all x, the sequence $g(x, 0), g(x, 1), \ldots$ is ultimately 1 or 0, according to whether $x \in S$ or $x \notin S$ [...] [I]t is shown that [...] Limiting recursive is equivalent to 2-recursive (EA and AE). (Gold, 1965, p. 28)

In slightly more modern times, we have e.g., (McCarthy and Shapiro, 1987) where, again, the same idea is given the name *Turing projectable* sets. Here, a generalized "extended Turing machine", a model of *non-terminating*, but effective, computational processes is described. This is a deterministic machine with two tapes: the projection tape and the computation tape. A computation is a finite sequence of configurations of this machine, and the output of a computation is (the number described by) the contents of the projection tape, but unlike ordinary Turing machines, it is *not* necessary for the last configuration to be one in which the machine has halted. A computation is *stable* if extending it does not change the output, and a machine M is said to *project* a number-theoretic function f if, for each n, there is a computation of M which has the stable output $f(n)$. The fundamental result is:

> **Theorem** [...] A number-theoretic function is Turing projectable if and only if it is recursive relative to the halting problem for ordinary Turing machines. (McCarthy and Shapiro, 1987, p. 523)

Putnam's theorem on trial and error predicates and Δ^0_2 sets then follows as an easy corollary. The rest of the paper is devoted to applications to problems concerning learning strategies and so-called inductive logic.

Moving on to Jeroslow's work, he too had contemporaries with very similar ideas. At roughly the same time, R. Magari and C. Bernardi were working on what they called *dialectic systems* (and associated dialectic sets), published in (Magari, 1974). and there was also some collaboration between these projects. Magari's presentation is rather more technically involved

than Jeroslow's, but, roughly, a dialectic system is a triple (h, f, c) where h, f are total recursive, f is a permutation of ω, and $c \in \omega$. Furthermore there are requirements on h that the image of a set including c is the whole of ω, and that the image of any set is non-empty. If we consider the natural numbers as representing formulas in a logical language via a recursive coding, we may think of h as representing a deductive formal system, f as representing a mechanical method of generating (non-logical) axioms, and c as an arbitrary contradiction. Magari then defines the set of *theorems* of a system (h, f, c) as the limit of a revisionary process where, basically, the "axioms" $f(n)$ are provisionally added, but subject to later removal should it turn out that c becomes "derivable". A set is dialectic if it is the set of theorems of such a system.[19]

Recently, a project has been initiated to build upon and refine Magari's work, presented in the paper (Amidei et al, 2016a).[20] Here the authors generalize the dialectic systems to "quasidialectical systems", with the philosophical aim of having a formalism more in tune with a serious empiricist position in the philosophy of mathematics. To the systems of Magari are added a c^-, encoding *other reasons* for revising a theory, than flat out contradiction, and a function f^-, with the task of *replacing* an abandoned axiom in some computable manner. This gives more descriptive power, i.e., there are quasidialectical sets which are not dialectic sets in Magari's sense, but not the other way round, and on the philosophical side, we may get a possibility to formally represent a revisionary dynamics in mathematical theory building in accordance with mathematical practices.[21]

Ending this section, I would like to draw attention to two recent papers, with very different perspectives and methods, but both of considerable interest to anyone working in this general area.

[19] The dialectic sets form, unlike the experimental logics, a proper subclass of the Δ^0_2-sets.

[20] A second, more technical, part of the paper has recently been published as (Amidei et al, 2016b).

[21] A suggestion pointing in this general direction can be found in (Jeroslow, 1975, p. 255): "To proceed in that direction, more would have to be added to experimental logics. In addition to the body of knowledge currently asserted, which is represented by the current theorems [...] one would wish to make explicit experimentation with other assertions currently being viewed as being of various degrees of likelihood. I.e., one would wish to spell out the trial-and-error activity with non-axioms which are being screened for potential axiomhood."

M. Mostowski has published several articles on *FM-representability*, e.g., (Mostowski, 2008). This concept captures the property of an arithmetical relation being such that there is a formula representing each finite part of its characteristic function in all sufficiently large finite initial fragments of a standard model of the natural numbers. A beautiful chain of representability results are proved, from which we learn that to the previously known equivalences—being Δ_2^0, being a trial and error relation, being of Turing degree $\leq 0'$, etc.— we can add *being FM-representable*, and others more besides, among them *being statistically representable*, and *being decidable by a Zeno machine*.

With a clear cognitive science outlook, the survey paper (Isaac et al, 2014) starts from a methodological assumption referred to as a psychological Church-Turing thesis: "The human mind can only solve computable problems." The paper goes through a variety of applications of logic in general, and concepts from computational complexity theory in particular, to cognitive and experimental psychology. Due to the task-oriented nature of the latter, and the tendency to think of the basic activity of cognitive agents as information processing, this computational perspective promises to be of great utility:

> The value of the computational perspective is in its fruitfulness as a research program: formal analysis of an information processing task generates empirical predictions, and breakdowns in these predictions motivate revisions in the formal theory. […] [A]ll logical models of cognitive behavior (temporal reasoning, learning, mathematical problem solving, etc.) can strengthen their relevance for empirical methods by embracing complexity analysis […] (Isaac et al, 2014, pp. 818f)

Many diverse, but related, applications are presented, such as typical tasks from (cognitive) experimental psychology, non-monotonic reasoning, neural network implementations, semantic automata and even some comments on social cognition.[22]

[22]A book written with the same general perspective on the interplay between formal logic and empirical science about reasoning is (Stenning and van Lambalgen, 2012).

2.2 Convergence in science and its language

As indicated already in Section 1.1, the semantic part of the thesis was not created ex nihilo, but rather came into existence by letting pragmatist ideas on convergence interact with ideas on the meaning of terms which are applied differently over time. This section spends a few more words on spelling out the background, with the aim of setting a philosophical stage for the results to come. There are no technicalities whatsoever here, in stark contrast with Section 3.2, for which the following is a preparation.

2.2.1 Peirce and the origins of pragmaticism

The idea of scientific convergence, and its connection to theories of meaning, of course goes way back in the history of (modern) philosophy. One *locus classicus* is certainly C. S. Peirce's *How to Make Our Ideas Clear* (1878), which I here will use to introduce some tenets of pragmaticism.

In the context of making critical comments about older conceptions (especially Descartes') of "clearness" of ideas, Peirce introduces his three grades of clearness of apprehension. The first one is, roughly, mere familiarity with an idea, while the second is the access to a suitable definition. Now, these conceptions are present in older philosophy, but what he finds lacking is a third grade, which he associates with what has become dubbed the "pragmatic maxim".

Peirce seems to present a distinctly operational concept when he explains how we are to understand the term "belief".

> [A belief] has just three properties: First, it is something that we are aware of; second, it appeases the irritation of doubt; and third, it involves the establishment in our nature of a rule of action, or say for short, a *habit*. (Peirce, 1878, p. 129)

He later adds that:

> [W]hat a thing means is simply what habits it involves. Now, the identity of a habit depends on how it might lead us to act, not merely under such circumstances as are likely to arise, but

> under such as might possibly occur, no matter how improbable they may be. […] [T]here is no distinction in meaning so fine as to consist in anything but a possible difference of practice. […] It appears, then, that the rule for attaining the third grade of clearness of apprehension is as follows: Consider what effects, which might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of these effects is the whole of our conception of the object. (Peirce, 1878, pp. 131f)

In the final section of his paper, Peirce applies his "rule" to the fundamental concept of *reality*. Taking familiarity as unproblematic, the second degree can still be puzzling to most, says Peirce, but also suggests that philosophical analysis has come up with a workable definition by contrasting reality and fiction, so we may "define the real as that whose characters are independent of what anybody may think them to be" (Peirce, 1878, p. 137). But, importantly, this is still not *perfectly* clear; in fact Peirce comments that it would be a "great mistake" to suppose it is.

In accordance with the pragmatic maxim, the correct way of analyzing reality is to see to the sensible effects it involves, and, says Peirce, this means looking at beliefs, because that is what is relevant here: reality's power to cause beliefs. So the real philosophical problem is to distinguish *true* belief, i.e., belief in real things, from *false* belief, i.e., belief in the fictitious. The year before, Peirce had in his (1877) scrutinized four different methods of belief fixation and had come to the conclusion that only the scientific method could in the long run be successful. And when he uses the word "true" here, it is in his opinion a predicate only properly applied in a scientific context. So when is a belief true, then?

> [A]ll the followers of science are fully persuaded that the processes of investigation, if only pushed far enough, will give one certain solution to every question to which they can be applied. […] They may at first obtain different results, but, as each perfects his method and his processes, the results will move steadily together toward a destined centre. […]

> No modification of the point of view taken, no selection of other facts for study, no natural bent of mind even, can enable a man to escape the predestinate opinion. [...] This great law is embodied in the conception of truth and reality. The opinion which is fated to be ultimately agreed to by all who investigate, is what we mean by the truth, and the object represented in this opinion is the real. (Peirce, 1878, pp. 138f)

Peirce seems to be quite convinced that we are really bound to end up at a particular place, and that this is independent of our particular interests and other mental faculties. Reality is that thing which keeps us in check; we do not construct the world, as it were. On the other hand, this convergence may not happen any time soon, and there is *some* kind of dependence on "mind".

> [R]eality is independent, not necessarily of thought in general, but only of what you or I or any finite number of men may think about it [...] Our perversity and that of others may indefinitely postpone the settlement of opinion; it might even conceivably cause an arbitrary proposition to be universally accepted as long as the human race should last. (Peirce, 1878, p. 139)

This is, in a nutshell, Peirce's early pragmaticism. One thing conspicuously lacking from this account is any real discussion of *language*. Perhaps one could add a rather important property of beliefs to Peirce's list of three; a belief can typically be shared, or communicated using a declarative sentence. A pair of questions which almost force themselves upon us is then to what extent linguistic meaning is dependent on "what we think", and how to handle the fact that not only our belief sets, but the very meaning of the terms we use to express beliefs, seem to vary over time. This leads naturally over to the next section.

2.2.2 Putnam on the meaning of 'meaning'

In his highly influential (1975) paper, Putnam is making a case for treating natural-kind words as "indexical", or rigid designators in Kripke's parlance.

He states that traditional theories of meaning rested on two, mostly unchallenged, assumptions.

1. Knowing the meaning of a term is just a matter of being in a certain psychological state.

2. The meaning of a term (in the sense of "intension") determines its extension (in the sense that sameness of intension entails sameness of extension).

Putnam sets out to show that "these two assumptions are not jointly satisfied by *any* notion, let alone any notion of meaning." (Putnam, 1975, p. 136) So, according to (1) and (2), if two terms differ in extension, they differ in meaning, and since knowing the meaning of the terms consists in being in (two different) psychological states, these states actually determine the extensions (by determining the intensions).

The rest of Putnam's paper is mostly devoted to detailing a host of examples and thought experiments, which have become well-known denizens of the philosophical landscape, using Twin Earths, water-like substance which is "XYZ" rather than H_2O, Martian tigers, machine-lemons, pencil-organisms and whatnot. Their main function is to drive the point home that:

> [...] it is possible for two speakers to be in exactly the same psychological state (in the narrow sense), even though the extension of the term A in the idiolect of the one is different from the extension of the term A in the idiolect of the other. Extension is *not* determined by the psychological state. (Putnam, 1975, p. 139)

A central idea of Putnam's, and an, at least partial, explanation of failures of assumptions (1) and (2), is the principle of *division of linguistic labour*. When I, a non-expert in metallurgy, acquire a term like "molybdenum", I do *not* (in general) acquire any "concept" that fixes the extension. In particular, I do not have to be in possession of a method for discriminating between examples and non-examples (of molybdenum). To the extent that such methods exist, they are present in the linguistic community as a

whole—not in an individual mind.[23] When I, as an average speaker, use the word "molybdenum", the community sees to it that I manage to speak of something determinate. And the community does not do this by itself, it needs help from *reality*; this is not a theory of social constructivism. The very existence of the natural kind which is rigidly picked out by the term, seems to be a necessary condition.

After his first few, science fiction flavoured, examples, Putnam proceeds to discuss meaning of scientific (natural kind) terms over time, explicitly arguing against a certain kind of anti-realist position. He urges us not to conflate the meaning of a term and the *criteria* (methods, observations, theories) we, at some point in time, happen to be using to demarcate the term's extension. In this example, we are to imagine that there are pieces of metal which could not have been determined *not* to be gold by the methods available to Archimedes in his time, but which, by the operational criteria we have at our disposal *today* are seen not to be gold. Putnam's claim is that:

> […] "gold" has not changed its *extension* (or not changed it significantly) in two thousand years. Our methods of *identifying* gold has grown incredibly sophisticated. […] Archimedes would have said that our hypothetical piece of metal X was gold, but he would have been *wrong*. But *who's to say* he would have been wrong? The obvious answer is: *we are* (using the best theory available today). (Putnam, 1975, p. 153)

With respect to the theme of this thesis, the central part of the meaning theory presented by Putnam is that there are these two aspects working in conjunction: (i) meanings are social; and (ii) meanings are indexical. An important difference between, say, proper names, and terms like "gold" is that we can know and use a proper name to refer to an individual without knowing anything about said individual. When it comes to "natural kind terms", on the other hand, we are required to know *something* about stereotypical representatives of the kind, we have to have some individual, mental, conception.[24]

[23] Non-linguistic parts of the community can also be important, of course.
[24] Not reference fixing in itself, of course.

> We have now seen that the extension of a term is not fixed by a concept that the individual speaker has in his head, and this is true both because extension is, in general, determined *socially*—there is division of linguistic labor as much as of "real" labor—and because extension is, in part, determined *indexically*. The extension of our terms depends upon the actual nature of the particular things that serve as paradigms [...] (Putnam, 1975, p. 164)

The technical work presented in Section 3.2 is based on an outlook in philosophical semantics (and metaphysics) which, in a sense, accepts the social aspect, but denies, or at least does not want to rely on, the indexical aspect.

3 Contributions

Wherein the main points of the original research of the author's attached papers (Kaså, 2012, 2015, 2016, 2017) are presented.

3.1 Anti-anti-mechanism and experimental logics

In this section it is investigated whether Jeroslow's experimental logics, and their Δ_2^0-sets of theorems, can help us cut through the thorny discussions about the (ir)relevance of the so-called "limitative" theorems in metamathematics to questions about mechanistic models in the philosophy of mind.

3.1.1 Limitative theorems and anti-mechanism

There is a rather large set of papers and books in the logico-philosophical literature which address the question whether the human mind is—or could be—mechanical, and in particular whether even the arithmetical faculties can be represented as a Turing machine, or equivalent abstract device. The usual starting point for this discussion is (Lucas, 1961), where he famously puts forward his anti-mechanistic thesis, arguing that:

> Gödel's theorem states that in any consistent system which is strong enough to produce simple arithmetic there are formulae which cannot be proved-in-the-system, but which we can see to be true. (Lucas, 1961, p. 121)

And he goes on to say:

> 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.
>
> Gödel's theorem seems to me to prove that Mechanism is false, that is, that minds cannot be explained as machines.

The standard objection to Lucas's reasoning is that all we know for certain about such a system S is that *if* the system is in fact consistent then (e.g.) $S \nvdash \mathrm{Con}(S)$. But, of course, (i) there is no stopping *that* claim from being a theorem of S, and (ii) there is, in general, no reason to assume that "we" *know* that S is consistent, even though it is. As said above, a huge debate followed, and continues to this day, and it took some new turns when Penrose published his *Shadows of the Mind* (1994). We will not wade through the material here; to do justice to the whole debate would be to write another and quite different dissertation altogether. But for anyone interested in digging in, there is a selection of important works in the references of (Kaså, 2012).[25] Before I describe my own contribution, we will look at some additional background.

Lucas was definitely not the first wanting to draw, roughly, this kind of conclusions from the limitative theorems of Gödel and others. Gödel himself, in his "Gibbs lecture" in 1951, said that:

> [I]f the human mind were equivalent to a finite machine, then objective mathematics not only would be incompletable in the sense of not being contained in any well-defined axiomatic system, but moreover there would exist *absolutely* unsolvable diophantine problems [...] where the epithet "absolutely" means that they would be undecidable, not just within some particular axiomatic system, but by *any* mathematical proof the human mind can conceive. So the following disjunctive conclusion is inevitable: *Either* [...] *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* [...] It is this mathematically established fact which seems to me of great philosophical interest. (Gödel, 1951, p. 310)

Even this more careful disjunctive claim has received its fair share of criticism over the years. For one thing, it may not be so obvious that there even

[25] Some of this philosophical discussion is very enlightening, and some not so much. Here I would just like to mention two particularly worthwhile contributions: (Franzén, 2005) and (Lindström, 2006).

exists a well-defined set of "humanly solvable problems" (if one doesn't mean the presumably finite set of problems the human species will actually manage to solve before it goes extinct). But more than this, the "representational relation" between *the human mind* on the one hand and *finite machines* on the other is certainly not crystal clear.

In the introductory note to the published Gibbs lecture, Boolos shows sound philosophical sensitivity towards this issue, and writes:

> What may be found problematic in Gödel's judgement that his conclusion is of philosophical interest is that it is certainly not obvious what it means to say that the human mind, or even the mind of some one human being, *is* a finite machine, e.g., a Turing machine. And to say that the mind (at least in its theorem-proving aspect) or *a* mind, may be represented by a Turing machine is to leave it entirely open just *how* it is so represented.[26] (Boolos, 1995, p. 293)

3.1.2 The Hazen intervention

Hazen's take on this discussion appeared as two thought-provoking postings to the *Foundations of Mathematics* mailing-list, (Hazen, 2006a,b). He certainly chimed in with the majority, being unconvinced by the anti-mechanistic arguments of Lucas (and Penrose), but he wanted to present a new kind of counter-argument—a "positive" argument.[27]

Noting that there are many different "Lucas-Penrose theses" and "Lucas-Penrose arguments", he sets out what he calls his "favorite version" of Lucas's position.[28]

[26] Boolos also adds: "[T]he following statement about minds, replete with vagueness though it may be, would indeed seem to be a consequence of the second theorem: If there is a Turing machine whose output is the set of sentences expressing just those propositions that can be proved by a mind capable of understanding all propositions expressed by a sentence in class A, then there is a true proposition expressed by a sentence in class A that cannot be proved by that mind". Not nearly as exciting, but more reasonable than what Gödel says.

[27] From where I am standing, what Hazen does is elaborating Boolos's point about the representational relation.

[28] The wording has been changed in minor ways.

1. If the mind is mechanical, human mathematics is the product of a machine.

2. The product of a machine is a computably enumerable set, i.e., the set of theorems of some formal theory.

3. No arithmetically adequate, consistent, formal theory has a theorem asserting the consistency of that selfsame theory (Gödel's second incompleteness theorem).

4. But human mathematicians can know that their mathematics is (or will be) consistent.

5. Hence: The mind is not mechanical.

By giving an alternative account of what it can mean to be "the product of a machine", Hazen's plan is to undermine (2) in such a way that (4) is applicable to such a machine to the same extent as to the human mathematical mind. Then Gödel's incompleteness theorems would not be relevant, and the conclusion (5) would not follow.[29]

Looking at all mathematical statements that have been made over time, this set is blatantly inconsistent, so there is no way in which we can "know it to be consistent". So if (4) is to have any credibility, we have to take the parenthetical "or will be" remark seriously. One of the powers of the general human intellectual enterprise is our capacity to *proceed by trial-and-error*; this is our way to weed out mistakes and inconsistencies. The set of "humanly provable mathematics" cannot be everything we at some time proved using some system of axioms and rules, but which set is it?

Neither Hazen nor I have to invent a model for this. It is already there in the literature, and we have only to flip a few pages back to Section 2.1. Let us say that the real mathematical theorems, in the limit of our research, constitute a semi-Euclidean set. If, moreover, the evolution of our theories can in some sense be represented as being mechanical, then this set can be

[29]This is, of course, not to say that there couldn't be *other* convincing anti-mechanistic arguments.

identified with a set of theorems of a convergent experimental logic.[30]

If so, premise (2) of Hazen's version of Lucas's argument is definitely out, since we moved from Σ^0_1 to Δ^0_2, and we just need to make it probable that (4) can be applicable to such a machine. But this, as the reader surely already guessed, is just Jeroslow's Theorem 3 in 2.1.1.

So this is Hazen's case. There seems to be a quite legitimate view according to which a Δ^0_2-set is "the product of a machine", and Jeroslow's experimental logics adhere to (a mechanical version of) a semi-Euclidean picture of human mathematical activity. Furthermore, Gödel's second incompleteness theorem only survives in a relativized form, and self-consistency can be provable, in the new sense of provable, i.e., being a recurrent (and also stable) formula. All well? Not really.

3.1.3 A proper experimental logic

First: Boolos, Shapiro and Hazen have me convinced that the relation between the representation and the represented is not as straightforward as that between a Turing machine and a formal system. And I do think that Jeroslow's model has great merit, and that something like this helps us better understand what is going on in the discussion. But there are reasons to feel somewhat unsatisfied by the reasoning as presented.

Consider the proof that there are experimental logics proving their own consistency. This is not new; as we know, it is just an adaptation of a proof of Feferman about ordinary theories, but with non-standard proof predicates. So one gut reaction could be that "we don't usually say that Feferman (contra the usual formulation of the second incompleteness theorem) has proved that PA *can* prove its own consistency, after all, so why would we say so in this Hazen/Jeroslow case?" Such an immediate reaction would be misguided, though, since the consistency statement, *viz.* $\neg\forall s\exists t > s\exists p\,\eta(t, p, \bot)$, is indeed intensionally correct; it expresses the

[30] This is, of course, a highly contentious assumption. Jeroslow has a discussion about this, and says, inter alia, that it is at least "consistent with a certain positivist view, that of a mechanistic creature in a mechanistic universe. [...] The completely mechanical nature of the experimental logics is not objectionable in this setting, since it is hard to see what physical events could influence views on the proper axioms for mathematical objects, beyond the results of computations, which are mechanical." (Jeroslow, 1975, pp. 254f)

right property. But there is still something amiss. Look at the set $\mathrm{Th}(H)$ in the proof. It is anything but dynamic. In fact, it is just the deductive closure of $T + \mathrm{Con}_T$ at every point in time. So this is a Σ^0_1 set, and therefore arguably not a particularly suitable example for Hazen to use.

What one would like to see is a *proper* experimental logic, i.e., a definition of an H such that $\mathrm{Th}(H) \in \Delta^0_2 \setminus (\Sigma^0_1 \cup \Pi^0_1)$. And moreover, one would like to see an example of such an H which is self-supporting, in the sense of having its own canonical consistency statement in $\mathrm{Th}(H)$. This is what was done in the technical part of (Kaså, 2012).

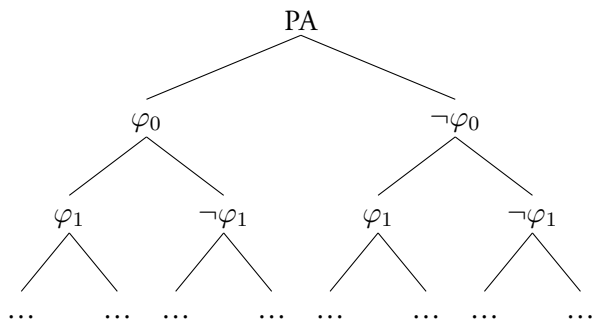
A *recursive Lindenbaum completion* of the decidable set of axioms PA, is defined over a *computable* enumeration $\varphi_0, \varphi_1, \varphi_2, \ldots$ of all closed formulas in the arithmetical language, like this.

$$\begin{cases} S_0 & = & \mathrm{PA} \\ S_{n+1} & = & S_n + \varphi_n, \text{ if this is consistent, otherwise } S_n + \neg\varphi_n \\ S & = & \cup S_n : n \in \omega \end{cases}$$

We immediately observe that:

- $S \notin \Sigma^0_1$ by Gödel-Rosser-Craig since it is a consistent, negation complete extension of PA;
- $S \notin \Pi^0_1$ since it is consistent and being a deductively closed extension of PA it is complete for true Σ^0_1-sentences.

More graphically, what we are doing when we are "completing", is to take a path down the infinite binary tree:



In general it is of course not a decidable problem to "check for inconsistency" at a node in the tree, but it is actually easy to slightly tweak the process, and get a convergent experimental logic H such that $\mathrm{Th}(H) = S$, i.e., to present S in a semi-Euclidean manner. The basic trick is to only consider small proofs of inconsistency (bounded in size by which step we are at in the construction) at each point, whose existence *is* decidable. This will lead to a revisionary process, so that sentences that were added at a previous stage may be later removed, put back in again and so forth, and we get an inductively defined sequence of sets $(T_0, T_1, T_2, \ldots)$. To get some "stabilization" over time, it turns out to be sufficient to prioritize early formulas over later ones when choosing what to get rid of, and in the end, the experimental logic can be defined by $H(t, p, \varphi) \Leftrightarrow (\varphi \in T_t \wedge p = \varphi)$. For details, see (Kaså, 2012, pp. 220f).

The next step is to tweak this construction to get a proper experimental logic, which is also self-supporting. We use a fixed-point argument, and briefly, the proof is this.[31]

1. An experimental logic of the type considered above, i.e., a recursive Lindenbaum completion, is determined by the effective enumeration.

2. Given a binary formula $\xi(x, y)$ (which might binumerate such an enumerating function) we define $\mathrm{Con}(\xi)$ as the canonical consistency statement for the corresponding experimental logic.

3. Take any computable enumeration $m(n)$ of all closed formulas in the arithmetical language, and consider:

$$F(n, x) = \begin{cases} \mathrm{Con}(x) & \text{and} \quad n = 0 \text{ ; or} \\ m(n-1) & \text{and} \quad n > 0. \end{cases}$$

4. The Recursion theorem gives a fixed point $\{\xi\}(n) = F(n, \xi)$.

[31] See (Kaså, 2012, pp. 221f). Note, however, that the paper uses the Fixed-point lemma for arithmetical theories, while here the Recursion theorem does the job. The basic idea is obviously the same in both cases.

5. This ξ gives a computable enumeration, and thereby an experimental logic of the kind discussed.

6. $\mathrm{Con}(\xi)$ is true, so it is certainly consistent with PA, and being first in the enumeration it is immediately added in the revisionary process of defining the set, and thus stays in.

So we have a strengthening of Jeroslow's result:

Theorem 5. *There are proper experimental logics which have their own canonical consistency statements as theorems.*

3.1.4 Final remarks on knowable consistency

So far, my contributions have been helpful to Hazen. But I still wish to express some doubts about the alleged impact of experimental logics on the question of knowable self-consistency.

Basically, I am just not sure that the perspective is right. What, if indeed anything, could it possibly mean to say that "a system (of the semi-Euclidean kind) knows that it is consistent"? Let us for simplicity say that *knows* just means *contains as a recurrent formula*, and suppose that "H is consistent" is in $\mathrm{Th}(H)$ in the form $\mathrm{Con}(\eta)$.[32] It is natural for us "on the outside" to say that $\mathrm{Con}(\eta)$ expresses the consistency of the experimental logic in question, and that it does so in an intensionally correct way, but our vantage point does not seem to be the relevant one. How does it look from "within"? Well, if only the system knew which system it is then it would certainly have knowledge of self-consistency. But, alas, this does not seem to be the case. $\mathrm{Th}(H)$ would have to contain an arithmetical statement expressing something along the lines of "I am H" (or rather "I am η"), and no such formula is forthcoming.

Furthermore, to be fair to Lucas, we have to take the relativized incompleteness results into account. As Jeroslow remarks in (Jeroslow, 1975, p. 264) this means that it "cannot prove its soundness for certain trial-and-error statements". In a similar vein, Theorem 5 in (Bennet, 1989) shows

[32] For some formula η defining H.

that we can *effectively* find an undecidable Π^0_2-sentence.[33] These relativizations of the incompleteness phenomena to the experimental case seem to give Lucas some room to maneuver. He can claim that we have exposed a deficiency in the machines that we humans do not suffer from, much as he did before. And even if the undecidable sentence mentioned above does not belong to "real mathematics" (being of too high complexity) it must "be accepted as meaningful if the concepts of experimental logics are so accepted." (Jeroslow, 1975, p. 264)

Note also that we in any case have the incompleteness issue.

> Under the mechanist view, the least non-trivial goal that can be set for human knowledge, must be that of obtaining as many true predictions $(\forall x)R(x)$ as possible [...] Our extension of the First Incompleteness Theorem [...] shows that the simultaneous requirements on an experimental logic, of consistency, convergence, and closure under reasoning, is inconsistent with the goal of obtaining *all* true predictions $(\forall x)R(x)$.
> (Jeroslow, 1975, p. 257)

But, on the other hand, as Boolos remarks while commenting on Gödel's "disjunctive result":

> [A] further problem for Gödel's view is that the supposition that the second alternative holds does not seem particularly surprising or remarkable at present. [...] Why, we may wonder, should there *not* be mathematical truths that cannot be given any proof that human minds can comprehend?
> (Boolos, 1995, pp. 293f)

I have to say that I wholeheartedly agree with this sentiment.

3.2 Logic sub specie aeternitatis

We now turn to the semantic part of my contributions, and a presentation of the sought-after formal explications of the terms *convergent concept* and

[33] A sentence φ such that PA $\vdash \forall t(H(t,p,\varphi) \rightarrow \exists s > t H(s,p,\neg\varphi))$. With some extra assumptions we can get it as Δ^0_2, provably in H.

true-in-the-long-run. The philosophical point of departure will be a "selective fusion" of ideas from Peirce and Putnam, tempered by Waismann (1945), to ground the explicatory work. The goal is an explicatum which represents the point of view of eternity, as it were. As for the technical results (and this section is mostly technical) the principle is the same as in the previous section; with few exceptions, the reader is referred to the papers for full proofs of lemmas and theorems.

3.2.1 Philosophical motivation

In Section 2.2.2, we saw that Putnam seems to have compelling arguments to the effect that we should not identify the meaning of a term in (empirical) science with its corresponding contemporary classificatory criteria, or methods for identification. Neither should we identify it with any kind of psychological concept. Instead the meaning-relation between a term and that which it signifies is upheld by a complex social structure, with sometimes sharply divided linguistic (and other) labour, and it is also underpinned by an ontology of natural kinds, which we manage to point to rigidly. Now, I have earlier in this text expressed sympathy towards some kind of Peircean (or Ramseyan) pragmaticism, as well as skepticism when it comes to natural kind metaphysics. One may then wonder what to take home from Putnam, and what to abandon. I suggest that the key concept, to be used to extract a "still social and convergent, but pragmatist" theory of meaning from Putnam's paper, is *open texture*.[34]

Waismann, in (1945), argues that "most of our empirical concepts are not delimited in all possible directions", and this goes also for concepts in the sciences.

> The notion of gold seems to be defined with absolute precision, say by the spectrum of gold with its characteristic lines. Now what would you say if a substance was discovered that looked like gold, satisfied all the chemical tests for gold, whilst it emitted a new sort of radiation? "But such things do not happen." Quite so; but they *might* happen, and that is enough

[34]Or *porosity*, "Porosität der Begriffe", as it was originally called in German.

> to show that we can never exclude altogether the possibility of some unforeseen situation arising in which we shall have to modify our definition. [...] In short, it is not possible to define a concept like gold with absolute precision, *i.e.* in such a way that every nook and cranny is blocked against entry of doubt. That is what is meant by the open texture of a concept. (Waismann, 1945, pp. 122f)

This is not to say that the concept is *vague*, since "vagueness can be remedied by giving more accurate rules, open texture cannot [...] definitions of open terms are *always* corrigible or emendable".[35] (p. 123) Admittedly, this is not immediately a criticism of Putnam's ideas. One possible take on the matter is that this just further explains why we should never identify the meaning of a term with the "definition" we happen to be using at the moment. But that would be to miss an important part of Waismann's message, in my opinion.[36]

It is not just that we are in practice unable to pin down *gold* with a description that fixes the extension in every possible situation; the real point is that, like it or not, our language—even our scientific language—in principle works in such a way as to be open for more or less radical change over time. Waismann calls this the "essential incompleteness of an empirical description":

> Every description stretches, as it were, into a horizon of open possibilities: how far I go, I shall always carry this horizon with me. [...] [W]e can never eliminate the possibility of some unforeseen factor emerging [...] the process of defining and

[35]One might object, and have the objection sustained, that we certainly in principle are at liberty to just *stipulate* a meaning for "gold" once and for all, and thus making it non-open, much as we could eliminate vagueness. But one of Waismann's points is that empirical concepts typically don't work that way.

[36]Just to be scholarly clear here: Waismann's paper was published 30 years before Putnam's, so he is not in any way commenting on a social-indexical meaning theory of a Putnam-Kripke variety, but is instead discussing a possible criticism of verificationalist empiricism. Nevertheless, his powerful notion of porosity is free to use for our present purpose, of course. Whether Putnam ever considered Waismann's ideas in connection with his own theory of meaning, I don't know.

> refining an idea will go on without ever reaching a final stage. (Waismann, 1945, pp. 124f)

The way I read this, and the way to reconnect with pragmaticism, is that our way of scientifically, and linguistically, organizing our experiential material, may eternally be subject to change. It is not that I deny that reality is there and constrains us, but it does not in and of itself force extensions on our terms. Indeed, Waismann comments that:

> [T]here is always the chance that something unforeseen may occur [...] (*a*) that I should get acquainted with some totally new experience such as at present, I cannot even imagine; (*b*) that some new discovery was made which would affect our whole interpretation of certain facts [...] the data of observation are connected in a new and unforeseen way, that, as it were, new lines can now be traced through the field of experience. (Waismann, 1945, p. 127)

Does this mean that there is no convergence in (scientific) language?[37] Certainly not. Typical terms that stand the test of time and are useful for classification, prediction, systematization, etc., may in fact generally be convergent in some sense, but this is dependent on many factors, of which regularities of our external world is but one. We have the whole web of *other* concepts as well as our theories—both of which are evolving—and also such things as the sophistication of our scientific equipment, limitations in our mental capacity and what have you. All these things, working in conjunction, will determine if, and how, there will be convergence when it comes to the extensions of terms in our language. In a word: convergence does not have to come from rigidly pointing to natural kinds.[38]

Given these considerations, and the earlier discussion in the background chapter, the following is a reasonable, if a bit rough, statement of my position on how a pragmatist construal of concept convergence ought to look.

[37] Which would go against Peirce, Putnam, Jeroslow, and general, often unarticulated, basic metaphysical beliefs of scientific communities.

[38] An illuminating real-life example of meaning litigation can be found in Chapter 2 of P. Ludlow's *Living Words*, where he gives a short, but instructive, account of the debate over recent changes in application of the term "planet". (Ludlow, 2014, pp. 41–51)

- Conceptual convergence (probably) exists, but we cannot in general *know* that the application of a certain term will stabilize over time, let alone know that the contemporary classificatory criteria pick out the "correct" extension, so that we have already stabilized.

- In fact, if there is convergence, it is in general point-wise (or better: *entity-wise*) meaning that the "real" or "final" extension may be found only in the limit of scientific development over time. It may well be the case that there is never a finalized operational definition, even though the subset of entities for which the classification is final grows over time.

- Not all concept terms in scientific discourse need be convergent in this sense, but perhaps only a subclass of core concepts.

So I don't deny that there is, or at least can be, a proper extension (and not just contemporary criteria) of a scientific term like "gold", *viz.* the limit extension, but this may not be attained after any finite amount of scientific progress (and may not be recognized, even if attained). Furthermore, while we give up the basis in natural kinds in favour of a more instrumental view, we also need something new to be able to define *true-in-the-long-run* in a way which respects ontological commitment of our scientific theorizing.[39] One can distinguish three aspects of meaning-as-extension of a convergent scientific term, where Putnam acknowledges two: (i) the extension determined by the contemporary criteria used at a particular point in time; and (ii) the true limit extension. Our semantics also needs access to (iii) the actual infinite sequence of contemporary extensions of the convergent terms—not only the final, limit extension. Imagine a case where our best theories over time always entail there being something satisfying a certain concept term, while there is no particular thing which, in the long run, needs to be thus classified. It seems reasonable to say that we as a scientific community are then committed to the non-emptiness of this concept.[40] Note that it could even be the case that *every individual entity* in the long

[39]This is "ontological commitment" somewhat in the sense of (Quine, 1948).
[40]This is called *sub specie aeternitatis* meaning in Kaså (2015).

run will be classified as *not* falling under this concept, so that the limit extension of the term is empty. In a short slogan, the logic presented below is a semantics for this *tripartite meaning of 'meaning'*.

Putnam's trial and error predicates have been mentioned above, in connection with Jeroslow's experimental logics, and they indeed play a conceptual role here too. Recall that these constitute a generalization of the class of decidable relations, and are the relations R on natural numbers such that there exists a computable function f satisfying, for all $x_1, \ldots, x_n$:

$$\begin{cases} R(x_1, \ldots, x_n) & \Leftrightarrow & \lim_{y \to \infty} f(x_1, \ldots, x_n, y) = 1 \\ \neg R(x_1, \ldots, x_n) & \Leftrightarrow & \lim_{y \to \infty} f(x_1, \ldots, x_n, y) = 0 \end{cases}$$

In Putnam's own words, such relations are "decidable by 'empirical' means" (Putnam, 1965, p. 49). In the context of the present discussion, computability is not an issue—only entity-wise convergence over time, as mentioned above. So we can think of f, with a point in time plugged in for y, as corresponding to a time-bound operational definition, the totality of resources (instruments, theories, observations, etc.) used at a particular time to classify objects with respect to the concept term R. If we have this kind of convergence, we will (with some benign ambiguity) refer to R (or f) as a *trial-and-error classifier*.

3.2.2 Syntax and semantics

Syntactically, the language is just ordinary first-order logic, and the intuitive semantic ideas are the following.[41]

- The basic predicates of the language will denote trial-and-error classifiers.
- At each point in time, this language will be interpreted in the standard classical way, over a relational model.
- Existential sentences will get a meaning which respects ontological commitment (regardless of limit extension), while the truth of a universal sentence will depend on the limit extension.

[41] A first, undeveloped, sketch of this kind of semantics appeared in (Sahlin and Kaså Palmé, 2005, Sect. II).

Formally, the models of this semantics will be ω-sequences of classical models $(\mathcal{M}_0, \mathcal{M}_1, \ldots)$ such that:[42]

- all models in the sequence have the same domain and the same signature;
- for every tuple $(m_1, \ldots, m_n)$ of objects in the domain, and n-ary predicate P in the signature, either $\exists i \forall j > i : (m_1, \ldots m_n) \in P^{\mathcal{M}_j}$ or $\exists i \forall j > i : (m_1, \ldots m_n) \notin P^{\mathcal{M}_j}$;
- for each constant c in the signature, $\exists i \forall j > i : c^{\mathcal{M}_j} = c^{\mathcal{M}_i}$.

And the formal truth definition is:

Definition 1 (Trial-and-error truth)**.** $(\mathcal{M}_0, \mathcal{M}_1, \ldots) \vDash_{\text{te}} \varphi$ iff

(i) $\exists i \forall j > i : (c_1^{\mathcal{M}_j} = c_2^{\mathcal{M}_j})$, for $\varphi = (c_1 = c_2)$;

(ii) $\exists i \forall j > i : (c_1^{\mathcal{M}_j}, \ldots, c_n^{\mathcal{M}_j}) \in P^{\mathcal{M}_j}$, for $\varphi = Pc_1 \ldots c_n$;

(iii) $(\mathcal{M}_0, \mathcal{M}_1, \ldots) \nvDash_{\text{te}} \psi$, for $\varphi = \neg\psi$;

(iv) $(\mathcal{M}_0, \mathcal{M}_1, \ldots) \vDash_{\text{te}} \psi$ and $(\mathcal{M}_0, \mathcal{M}_1, \ldots) \vDash_{\text{te}} \gamma$, for $\varphi = (\psi \wedge \gamma)$;

(v) $\forall m \in \text{dom}(\mathcal{M}_0)\, \exists i \forall j > i\, (m \in \psi^{\mathcal{M}_j})$, for $\varphi = \forall x \psi$;

(vi) $\exists i \forall j > i\, \exists m \in \text{dom}(\mathcal{M}_0)\, (m \in \psi^{\mathcal{M}_j})$, for $\varphi = \exists x \psi$.

Note that we define trial-and-error *truth* rather than satisfaction, and that the two quantifier clauses are not done with typical Tarski-style recursion. The quantified sentences are interpreted by way of satisfaction for their subformulas (with at most x occurring freely), which is handled by the underlying *classical* semantics.

In accordance with the last bullet point above, from the intuitive basis for the semantics, universal quantification (v) is taken to mean that each object in the long run satisfies the concept in question (regardless of whether there will ever be an actual point where they all do so), while existential quantification (vi) tracks the ontological commitment in the long run.

[42] The symbolism $\sigma^{\mathcal{M}}$ is used for the denotation of σ in $\mathcal{M}$, as defined in the semantics for ordinary classical logic.

3.2.3 Some simple properties

It is immediate, by straightforward applications of the truth definition to suitable countermodels, that some valid sequents from classical logic are no longer so. Here are a few examples.[43]

1. $\forall x \neg Px \not\models_{\text{te}} \neg\exists x Px$ and $\exists x \neg Px \not\models_{\text{te}} \neg\forall x Px$.
2. There are formulas φ such that $\neg\exists x\varphi \not\models_{\text{te}} \forall x\neg\varphi$ and also φ such that $\neg\forall x\varphi \not\models_{\text{te}} \exists x\neg\varphi$.
3. Similarly, there are quantified φ such that $\not\models_{\text{te}} (\exists x\varphi \vee \exists x\neg\varphi)$.
4. $\exists x(Px \vee Qx) \not\models_{\text{te}} (\exists x Px \vee \exists x Qx)$
5. There are universally quantified φ such that $\varphi(c) \not\models_{\text{te}} \exists x\varphi$.

Examples such as these, in combination with the following results show that the trial-and-error logic is strictly weaker than classical logic.[44]

Lemma 1. *$\mathcal{M} \models_{\text{c}} \varphi$ if and only if $(\mathcal{M}, \mathcal{M}, \ldots) \models_{\text{te}} \varphi$.*

So, a "static" trial-and-error model has the same complete theory as its classical base. (Here and below, $\models_{\text{c}}$ is used to denote classical satisfaction.)

Corollary 1. $\Gamma \models_{\text{te}} \varphi \Rightarrow \Gamma \models_{\text{c}} \varphi$.

That is, trial-and-error logic is included in classical logic.

Looking at the semantics, we see that adding a "vacuous" quantifier ($\exists v$, say) in front of a formula which is already closed forces it to be almost everywhere true in the model sequence. This leads to this next result, again connecting the trial-and-error logic to ordinary logic.

Lemma 2. *$\Gamma \models_{\text{c}} \varphi$ if and only if $\{\exists v\gamma \mid \gamma \in \Gamma\} \models_{\text{te}} \exists v\varphi$*

A difference between classical logic and this trial-and-error variant is that we do not have general compactness, that is, a set of formulas may have consequences that finitary methods will not bring forth. But this result only holds for large vocabularies, as in seen in the final result of this subsection.

[43] See (Kaså, 2016, p. 7).
[44] (Kaså, 2016, p. 8)

Theorem 6. *There is a theory of cardinality* $> 2^{\aleph_0}$ *which is finitely satisfiable, but not satisfiable.*

This is proved by taking an index set I of cardinality $> 2^{\aleph_0}$, and considering $\{\neg\exists x P_\alpha x \mid \alpha \in I\} \cup \{\exists y(\exists x P_\alpha x \vee \exists x P_\beta x) \mid \alpha, \beta \in I, \alpha \neq \beta\}$, which can be used as a counterexample to compactness.[45]

While this theorem is technically interesting, the intended application of the semantics makes most sense for countable (or even finite) vocabularies, for which we will see in the next two sections that finitary reasoning is, indeed, sufficient.

3.2.4 Getting standard models

Before moving on to the proper theorems, we make a liberalizing definition, and then state a key lemma to the effect that this liberalization is innocent, at least in the countable case.

Definition 2 (General models)**.** Call a model sequence satisfying the convergence constraints (in Section 3.2.2) but of any linear order type without a right endpoint a *general model sequence*, and define trial-and-error truth over such sequences in exactly the same way as in Definition 1.

Given any countable general model, we form its "trial-and-error diagram", and then work through this theory in an systematic manner to extract (i.e., inductively define) an ω-sequence which trial-and-error satisfies the same theory. This proves the technically important result:

Lemma 3 (Omega lemma)**.** *For each general model* $(\mathcal{M})_I$ *with countable domain and countable signature* $\mathcal{L}$*, there is a proper trial-and-error model* $(\mathcal{N})_\omega$ *where the same* $\mathcal{L}$*-sentences are trial-and-error true.*

So, the natural thought that the series of events which are relevant for conceptual evolution form an ω-sequence is in a sense not particularly restrictive.

[45](Kaså, 2016, pp. 8f)

3.2.5 Translation and fundamental theorems

To prove that this trial-and-error logic enjoys some standard properties such as countable compactness and axiomatizability, we use "transfer" from classical first-order logic. A translation from trial-and-error logic into classical logic which tracks the truth conditions from the semantics is provided. This translation will capture trial-and-error logic over the general models mentioned in the previous section, which means that the omega lemma is exactly what is needed to get proofs of the sought-after theorems.

Essentially, starting with a set of sentences Γ in trial-and-error logic, the translation is effectuated by:

- introducing two new unary predicates for "times" and "objects in the trial-and-error domain";
- expressing, with a sentence Ord, that the "times" are linearly ordered without right endpoint;
- providing each non-logical symbol with an extra slot (to "plug in an argument for time");
- translating each formula $\gamma \in \Gamma$, to a γ^*, tracking truth conditions;
- stating, with a set Conv_Γ, the appropriate convergence constraint for each non-logical symbol occurring in Γ.

Then one can move between classical models of a set and a trial-and-error model of its translation, using the omega lemma when needed, to get the following result.

Lemma 4 (Translation)**.** $\Gamma \vDash_{\mathrm{te}} \varphi$ *if and only if* $\mathrm{Ord} + \mathrm{Conv}_\Gamma + \Gamma^* \vDash_{\mathrm{c}} \varphi^*$

This lemma paves the way for the above-mentioned transfer, and the fundamental theorems basically fall out as corollaries.

Theorem 7 (Compactness)**.** *Trial-and-error logic is compact for countable vocabularies.*

Theorem 8 (Downward Löwenheim-Skolem)**.** *If* Γ *is countable and trial-and-error satisfiable, then* Γ *has a model with countable domain.*

Theorem 9 (Undecidability)**.** *Trial-and-error logic is undecidable.*

Theorem 10 (Axiomatizability)**.** *The set of trial-and-error validities over a countable vocabulary is effectively enumerable.*

Proofs are in (Kaså, 2016).

3.2.6 The basic fragment and natural deduction

There is a natural fragment of the trial-and-error logic of the preceding sections, given by a restriction in the syntax to quantifier depth ≤ 1, but leaving the models and the truth definition intact. This is the way the logic was first published in (Kaså, 2015), where it also got independent motivation as a "logic for trial-and-error classifiers".

> From the human experimenters' point of view, 'x is a P' is considered true (now) if the classifier corresponding to P replies 'yes' when applied to the object x, a statement like 'all things are P' is true if all objects are thus potentially classified, and so on, and so forth. But this is not the perspective that concerns us in the present paper. What we want to do is to have a semantics suitable for "timeless descriptions". To use a perhaps overly picturesque language, we may say that the semantics ought to be suitable for a supreme, omniscient being, looking at our toil from a vantage point outside of time, but commenting on the expressions in our trial and error language. This being will have a good birds-eye view on our de facto commitments in the long run, but it seems reasonable to hypothesize that it will not care much about the details of what happens at each step in our intellectual development. [...] The syntactical restriction of the language to depth ≤ 1 is motivated by the objective described above [...] this paper treats only the tendentious concepts, and the external, timeless language is designed so that all quantification is over formulas corresponding to (finitary combinations of) trial and error classifiers. (Kaså, 2015, pp. 310ff)

With the depth restriction in place, the logic moves closer to classical logic (correspondingly restricted, of course). Familiar sequents such as $\neg\exists x\beta \Rightarrow \forall x\neg\beta$ and $\neg\forall x\beta \Rightarrow \exists x\neg\beta$ now come out as generally valid, and existential introduction is now sound. It is still a sub-classical logic, though, since many of the non-sequiturs persist, but it turns out that we can use a perfectly ordinary natural deduction system if we just add a new side condition for existential elimination.

$$\dfrac{\exists x\beta \qquad \begin{array}{c}[\beta(c/x)]\\ \mathscr{D}\\ \gamma\end{array}}{\gamma}\ (\text{te}\exists\text{E})$$

(i) c does not occur in γ, β or in any undischarged assumption of $\mathscr{D}$ except for $\beta(c/x)$. (Just like for $\exists$E in classical logic.)

(ii) Both γ and the undischarged assumptions in $\mathscr{D}$ are either quantifier-free or existential sentences.

Only slightly misleading, this can be put in words like this: Existential reasoning is "internal" in that it takes place inside a context which is itself existential (or Boolean).

This proof system is sound and complete with respect to the semantics (applied to the fragment), using roughly a Henkin-style proof, and falling back to the completeness of classical logic. The details are in the paper, but the most interesting parts are probably the following two lemmas. First, the argument for soundness of the revised existential elimination rule.

Lemma 5. te$\exists$E *is sound with respect to trial-and-error consequence.*

Proof. Consider the subderivation $\mathscr{D}$. It is also a sound derivation in classical logic, so we have $\{\delta_1, \ldots, \delta_n, \beta(c/x)\} \vDash_{\text{c}} \gamma$, where the δ_i are the undischarged assumptions, and since c is "arbitrary", $\{\delta_1, \ldots, \delta_n, \exists x\beta\} \vDash_{\text{c}} \gamma$. Now, if $(\mathcal{M}_0, \mathcal{M}_1, \ldots) \vDash_{\text{te}} \{\delta_1, \ldots, \delta_n, \exists x\beta\}$, the restriction in the te$\exists$E rule on δ_i gives that $\exists i\forall j > i : \mathcal{M}_j \vDash_{\text{c}} \{\delta_1, \ldots, \delta_n, \exists x\beta\}$, and therefore $\mathcal{M}_j \vDash_{\text{c}} \gamma$. By the restriction on γ, we have $(\mathcal{M}_0, \mathcal{M}_1, \ldots) \vDash_{\text{te}} \gamma$. □

Note how the special restriction is used. If, e.g., δ_i were universally quantified, it could well be trial-and-error true without being classically true in any model. And conversely, if the conclusion is $\gamma = (\exists x P(x) \vee \exists x Q(x))$, it could be classically true everywhere in the sequence without being trial-and-error true.

The next result allows us to pick classical models for certain trial-and-error consistent sets of sentences, which is used to take care of cofinally false existential statements in the final model construction in the proof of the completeness theorem.

Lemma 6. *If* $\exists x\alpha_1, \ldots, \exists x\alpha_n \vDash_{\mathrm{c}} \exists x\beta$ *then* $\exists x\alpha_1, \ldots, \exists x\alpha_n \vdash \exists x\beta$.

Proof. Assume $\{\exists x\alpha_1, \ldots, \exists x\alpha_n, \neg\exists x\beta\}$ is *not* classically satisfiable. This holds iff $\{\alpha_1(c_1/x), \ldots, \alpha_n(c_n/x)\} \cup \{\neg\beta(d/x) \mid d \in D\}$ is not classically satisfiable, where all c_i are new, and D is the set of all constants occurring in the formulas $\alpha_1(c_1/x), \ldots, \alpha_n(c_n/x), \beta$.[46]

By the completeness of classical logic, there exists a formal derivation $\alpha_1(c_1/x), \ldots \alpha_n(c_n/x) \vdash_{\mathrm{c}} \bigvee_{d\in D} \beta(d/x)$. Normalizing, we get a derivation using only propositional rules, which is therefore also a derivation in trial-and-error logic, from only quantifier-free assumptions.

Each disjunct $\beta(d/x)$ gives $\exists x\beta$ by $\exists$I, so after a suitable number of applications of $\vee$E, the result is a derivation, still in trial-and-error logic, $\alpha_1(c_1/x), \ldots \alpha_n(c_n/x) \vdash \exists x\beta$. This means that we can apply te$\exists$E n times, to sub-proofs with only quantifier-free and existential assumptions, to get a proof of $\exists x\alpha_1, \ldots, \exists x\alpha_n \vdash \exists x\beta$. □

Summing up what we learned from this (soundness and) completeness result, we may say that if we restrict ourselves to quantification over concepts which are guaranteed to be convergent, i.e., the trial-and-error classifiers, the difference between internal reasoning "inside the time-line" and reasoning from the external, and eternal, perspective comes down to restricted rules for existentials in the latter.

[46] We get this from a special case of Herbrand's theorem.

3.2.7 Analytic tableaux and decidability

In the last section we saw that the basic, depth-restricted, fragment of trial-and-error logic has a concrete natural deduction proof system. Unlike the full logic, which was easily seen to be undecidable, one would intuitively expect this fragment to be decidable, but it is obviously not possible to point to a "finite model property" in the ordinary sense, since it is quite easy to force models of even very simple sentences to have infinite domains. But there is a way of getting a decidability theorem, and, as a bonus, also a concrete algorithm for generating models for consistent sentences. This is through the introduction of a tableau proof system for the fragment logic, and also an alternative semantics, which has an effective finite model property, and can be proved to be, in the appropriate sense, equivalent to the standard semantics.[47]

We introduce something that is very much like an ordinary tableau system, but as was the case with natural deduction, we have to handle the peculiarities of existential reasoning. In this case this shows up in the introduction of *indices* for false existential statements (intuitively, to indicate at which points they are cofinally false in the model), and special *virtual parameters* to instantiate true existentials (intuitively, to be non-convergent witnessing functions).[48] The rules for the quantifiers are as follows:

Constant universal rules[49]

$$\begin{array}{c} t\forall vB \\ | \\ tB(a/v) \\ a \text{ is any} \\ \text{constant} \end{array} \qquad\qquad \begin{array}{c} [i]f\exists vB \\ | \\ fB(a/v) \\ a \text{ is any} \\ \text{constant} \end{array}$$

[47] There is no room here for introducing tableau systems in general. Any reader unfamiliar with such systems is urged to rectify that situation by consulting the excellent (Smullyan, 1995).

[48] Example tableaux can be found written out in full in (Kaså, 2017).

[49] I.e., "instantiated with constants".

Constant existential rule

$$\begin{array}{c} f\forall vB \\ | \\ fB(a/v) \\ \textit{new} \text{ constant } a \end{array}$$

Rules for virtual parameters

$$\begin{array}{c} [i]f\exists vB \\ | \\ fB(d(i)/v) \\ d(i) \text{ is any virtual} \\ \text{parameter with} \\ \text{matching index} \end{array} \qquad \begin{array}{c} t\exists vB \\ | \\ tB(d(i)/v) \\ d \text{ is } \textit{new} \\ i \text{ arbitrary} \end{array}$$

Next, we can give an alternative, and finitary, semantics based on the intuition that, while a trial-and-error model may be an infinite sequence of models with an infinite domain, the depth restriction for the fragment language means that we can describe the model sequence (for a finite set of sentences) in a finite manner.[50]

Definition 3. A *finitary trial-and-error model*, or simply *fte-model*, $\mathcal{M}$ is a finite sequence of classical first-order models $(\mathcal{M}_0, \ldots, \mathcal{M}_n)$ of the extended tableau language with a common *finite* domain and common interpretation of constants and virtual parameters. There is to be a designated non-empty subset of the domain, the *constant domain*, which, inter alia, interprets the constants, and these objects (and tuples of them) satisfy the same basic predicates in the whole sequence.

Definition 4 (fte-truth)**.** Here we define the relation $\vDash_{\text{fte}}$, i.e., what it means for a signed (and possibly indexed) formula in the tableau language to be true in an fte-model.

[50]This semantics interprets not only the original language, but also the signed and indexed formulas in the tableaux.

i) $\mathcal{M} \vDash_{\text{fte}} f\varphi$ iff $\mathcal{M} \nvDash_{\text{fte}} t\varphi$

ii) $\mathcal{M} \vDash_{\text{fte}} tB$ iff $\mathcal{M}_i \vDash_{\text{c}} B$ for all $0 \leq i \leq n$, if B is quantifier-free without virtual parameters.

iii) $\mathcal{M} \vDash_{\text{fte}} tB(d(i))$ iff $\mathcal{M}_i \vDash_{\text{c}} B(d(i))$.

iv) $\mathcal{M} \vDash_{\text{fte}} t\forall vA$ iff $\mathcal{M}_i \vDash_{\text{c}} A[m]$ for all m in the constant domain and all $0 \leq i \leq n$.

v) $\mathcal{M} \vDash_{\text{fte}} t\exists vA$ iff $\mathcal{M}_i \vDash_{\text{c}} \exists vA$ for all $0 \leq i \leq n$.

vi) $\mathcal{M} \vDash_{\text{fte}} [i]f\exists vA$ iff $\mathcal{M}_i \vDash_{\text{c}} \neg\exists vA$.

Setting things up like this, a proof of soundness and completeness for the tableau system with respect to this new semantics is not all that hard to come by.

Theorem 11. *The tableau system is sound and complete with respect to the fte-semantics.*

To show that the two semantics (te and fte) are equivalent, we need, in one direction, to pick out finitely many appropriate models and identify the constant domain, and in the other direction duplicate both the models and their domains to infinity. It turns out that there are simple algorithms for this, so the next result says that:

Theorem 12. *A depth-restricted set has an ordinary trial-and-error model if and only if it has an fte-model.*

As a direct consequence of the foregoing, and wrapping up this treatment of the ≤ 1-depth fragment of trial-and-error logic, we get this final theorem.

Theorem 13. *Satisfiability (and validity) for the depth-restricted logic is decidable, e.g., by using the proposed tableau system.*

Precise statements, and proofs, of the above results are found in (Kaså, 2017).

3.3 In conclusion

This work started with a conviction; that what *sorts of things* there are, or eventually turn out to be, is dependent not only on properties of the external world, but is as much a feature of *us*. What turns out to be useful classifications, in science and other rational endeavours, is at least partly determined by our human technical, cognitive and societal evolution. The world is there to keep us in check, to be sure, I am not a full-blown relativist (if such creatures exist) but rather some kind of a pragmatic instrumentalist.

I have not really argued for this position in the present book (though some arguments crept into Section 3.2.1); the mission has been of a different sort. Taking such an outlook as *given*, i.e., skepticism towards natural kinds, but belief in the prospect of some stability over time for our conceptual apparatus, the task has been to explicate the *convergent concepts*, introducing the trial-and-error classifiers, technically inspired by (Putnam, 1965), and building on philosophical ideas in (Peirce, 1877, 1878), (Putnam, 1975) and (Waismann, 1945).

While not in any sense completed (see the list of open problems below) the study so far has been rather rewarding, with insights into possible formal semantics for these classifiers and several important pieces of knowledge about which basic logical properties hold, and do not hold, for these trial-and-error logics. In the simple case where quantification is only over combinations of classifiers, we have seen that the differences in comparison to classical logic all come down to how reasoning with existential statements works (and I am particularly happy with the analytic tableau system). The general case is of course more intricate, but I have made some non-trivial headway there too. So I dare say that the *explicatum* has, without doubt, proved to be formally tractable (and quite interesting, even), but I do not pretend to have a definitive answer to whether it fulfills Carnap's criterion (3) on fruitfulness.

Being introduced to (Jeroslow, 1975) early in the process was surely a stroke of luck. Even though only obliquely connected to the semantic concerns, it proved to be inspirational indeed, outside its perhaps intended area of application. And what more is, the proposal in (Hazen, 2006a) intrigued me, and forced me to read up on the anti-mechanism debate, which was

in many ways enlightening. It definitely seems reasonable to problematize what it should mean that the mind can, or cannot, *be represented* by a machine (or by a formal system), and what the proper relation is between the "output" of the mind and the theorems of a system. Though not the last word on the subject, Jeroslow's experimental logics constitute, at the very least, a valuable entry point. In the process, I have gained some budding insight as to the possibility of representing trial-and-error processes in very general and abstract terms, and the role of complexity classes such as Δ_2^0 in that context. An answer to what it really is that makes certain formal consistency statements "natural" still eludes me, though.

But for the time being, my exploration is over, and it is time to sign off.

4 Some open problems

While this thesis work has now come to its inevitable end, the study of trial-and-error aspects of logic and semantics offers so many more interesting prospects. The ideas I have expounded can be further developed and extended, and there are surely related problems that I have not even thought of yet. By way of conclusion, I will just list what I find to be the most obvious, and promising, ideas for building upon what has been done so far. Anyone who would like to address these issues has my unconditional blessing.

1. It would be interesting to try to comprehensively describe the *philosophical* significance of Δ^0_2 (and subclasses thereof, and perhaps other complexity classes). Starting from literature mentioned above in Section 2.1.3, perhaps the most pressing point is to look into a realistic model of dynamic systems where propositions are vetted for acceptance as new axioms, and the system is expanded by trial-and-error. One should probably also connect this with work in the belief revision tradition.

2. A minor technical point, but one which I would like to look into myself, is devising further ways to tweak the simple construction of proper experimental logics (as in Section 3.1.3) to get sets of theorems with desired properties.

3. A question that almost asks itself, but to which I have so far not devoted much energy, is whether there are any technically useful connections between the theory-oriented Section 3.1 and the semantic Section 3.2 of this thesis.

4. There is a very natural sense in which trial-and-error logic can be seen as a proper part of a much more expressive quantified *modal*

language. If we have, e.g., axiom schemata for linear frames and add axioms like $(\Box\Diamond Px \to \Diamond\Box Px)$ for atomic formulas, we have something very much like a trial-and-error model. Then universal quantification becomes $\forall x\Diamond\Box\varphi$ and existential quantification $\Diamond\Box\exists x\varphi$. This connection should be explored.

5. One thing conspicuously missing from (Kaså, 2016) is a concrete proof system for the trial-and-error logic of that paper. It is shown to be axiomatizable (and compact for countable vocabularies), so it would be nice to provide a system which could deliver some insight into how trial-and-error *reasoning* really works, once the depth restriction is removed from the syntax.

6. Continuing from the previous item, one would also like to see some proper model theory, which could then provide more results than we get just from transfer from first-order logic. The first thing to look into is how to define *equivalence* between models in terms of a trial-and-error version of Ehrenfeucht-Fraïssé games.

7. There is a very natural, and prima facie much more powerful, extension of my trial-and-error logics, which may be dubbed "fully expressive". The idea is to change the syntax to allow formulas like $Q_1x_1 \dots Q_nx_n \mid \varphi$ with an arbitrary quantifier prefix and arbitrary classical first-order φ and then have basically the same kind of semantics, but with the trial-and-error prefix $Q_1x_1 \dots Q_nx_n$ now "picking objects from the domain" before the $\exists i\forall j > i$ part of the truth definition.

8. On the more non-technical side, I believe that it would be a worthwhile project to map out a comprehensive, philosophically coherent, pragmatist dynamic meaning theory, something really deserving the epithet "a real fusion of Peirce-Ramsey-Putnam-Waismann".

5 Brief summaries of the papers

5.1 Experimental Logics, Mechanism and Knowable Consistency

The point of departure of this paper is a suggestion by A. Hazen regarding a possible way to counter a certain kind of *anti-mechanist* arguments, *viz.*, arguments, based on Gödel's incompleteness theorems, to the effect that the mind cannot possibly be equated with a machine. Hazen's strategy is to abandon the preconception that if the human mind is a machine (explicated in Turing's sense), then the "output" of such a mind must be a computably enumerable set, and therefore equivalent to a set of theorems of a formal theory. These discussions are typically framed in the context of the mathematical (or even just arithmetical) faculties of the mind, and Hazen observes that an important part of mathematical activity is theoretical *revision*. The fact that we at some point "prove" something doesn't mean that it should count as a real theorem; we may well retract it later because our axiomatic base is open to revision. As a simple representation of such dynamic theories, Hazen employs the *experimental logics* of R. Jeroslow.

These systems are generalizations of ordinary formal theories, for which the concept of theoremhood is a Δ^0_2 property rather than Σ^0_1, and there are some basic theorems both extending and relativizing the incompleteness results.

In the paper these systems are presented, and a few results are somewhat sharpened, to try to substantiate Hazen's claims, which are also philosophically evaluated. Some semi-technical doubts are raised concerning the alleged impact of experimental logics on the question of knowable self-consistency.

5.2 A Logic for Trial and Error Classifiers

In the second paper we are introduced to the concept of a *trial-and-error classifier*, which is a formal explication of concepts (or terms in a language) which have an extension that changes over time, but exhibit at least a weak kind of convergence. The philosophical point of this is to capture the idea to be found in, e.g., classical pragmatism, that it is necessary for science, and arguably for rational communication in general, that while our concepts and our classifications may change, and perhaps change indefinitely, there is some kind of long term agreement.

A syntactical fragment of the language of classical first-order logic is given a new semantics, using ω-sequences of classical models, subject to some convergence constraints, so that the basic predicates can be interpreted as representing classifiers of this kind. This gives a formal meaning to claims of "being committed to a classification in the long run", and the corresponding logic makes it possible to ask questions of the type "If we are committed to all sentences in the set Γ, are we then also committed to φ?"

This logic is decidedly different from classical logic, but it turns out that we can use a natural deduction proof system which is almost standard, differing only when it comes to conditions for application of existential elimination. The paper contains a somewhat novel completeness proof for this formal system.

5.3 Formally Modelling Convergent Dynamic Meaning. Results on Compactness and Axiomatizability

This paper continues the study of classifiers. The syntactic restriction is now removed, and the trial-and-error semantics is applied to all formulas in the ordinary syntax of first-order logic.

After looking at examples illustrating some typical validities and non-validities for this semantics, and proving that the logic is non-compact for large vocabularies, the main part of the paper is spent on relating trial-and-error logic to classical logic via translation. This translation is designed so that it tracks the truth conditions from the semantics, and it can be shown that the translation of a set of formulas has a classical model if and

only if the set has a trial-and-error model in a generalized sense, where an arbitrary linear order is used instead of an ω-sequence. Then it is proved that, for countable vocabularies, it is always possible to extract a proper trial-and-error model from this generalized one, keeping the truth values of all formulas.

From these lemmas we get a transfer of desirable properties from classical logic to trial-and-error logic: downward Löwenheim-Skolem, countable compactness, and axiomatizability. It is also easy to see that the logic is undecidable.

5.4 Analytic Tableaux for Trial-and-Error Reasoning

Unlike the undecidable logic of Paper III, one would intuitively expect there to be an algorithm deciding satisfiability for the syntactic fragment presented in Paper II. At the same time, it is easy to see that the logic does not have a finite model property in the ordinary sense. The last paper introduces a new, *finitary*, semantics for this logic, which is proved to be, in the appropriate sense, equivalent to the semantics based on ω-sequences. This gives the expected decidability result.

There is also introduced an analytic proof system for the logic, i.e., a system where every step of the process just introduces subformulas of the preceding steps, in the form of so-called (semantic) tableaux. This tableau system is proved to be sound and complete with respect to the finitary semantics, and therefore, in the light of the equivalence, for the logic of Paper II. The syntactic restriction then implies that these tableaux can be used not only to prove validity, but also to prove satisfiability, in that the construction of a tableau gives a finite recipe for defining a model of any satisfiable set.

References

Amidei, J., Pianigiani, D., San Mauro, L., Simi, G., & Sorbi, A. (2016a). Trial and error mathematics I: Dialectical and quasidialectical systems. *Review of Symbolic Logic*, 9(2), 299-324.

Amidei, J., Pianigiani, D., San Mauro, L., & Sorbi, A. (2016b). Trial and error mathematics II: Dialectical and quasidialectical sets, their degrees, and their distribution within the class of limit sets. *Review of Symbolic Logic*, 9(4), 810-835.

Bennet, C. (1989). Incompleteness for experimental logics. In C. Åberg (Ed.) *Cum Grano Salis: Essays dedicated to Dick A. R. Haglund*. Acta Philosophica Gothoburgensia (Vol. 3, pp. 43–50).

Boolos, G. (1995). Introductory note to *1951. In *Kurt Gödel: Collected Works* (pp. 290–304).

Carnap, R. (1950). *Logical Foundations of Probability.* The University of Chicago Press.

Feferman, S. (1960). Arithmetization of metamathematics in a general setting. *Fundamenta Mathematicae*, 49, 35–92.

Franzén, T. (2005). *Gödel's Theorem: an Incomplete Guide to its Use and Abuse*. A K Peters, Ltd.

Gold, E. M. (1965). Limiting recursion. *The Journal of Symbolic Logic*, 30(1), 28–48.

Gödel, K. (1951). Some basic theorems on the foundations of mathematics and their implications. In *Kurt Gödel: Collected Works* (pp. 304–323).

Gödel, K. (1995). *Kurt Gödel: Collected Works* (Vol. 3). Oxford University Press.

Hazen, A. (2006a). Re: [FOM] The Lucas-Penrose thesis. www.cs.nyu.edu/pipermail/fom/2006-September/010809.html

Hazen, A. (2006b). [FOM] The Lucas-Penrose thesis (long). www.cs.nyu.edu/pipermail/fom/2006-October/010876.html

van Heijenoort, J. (Ed.) (1967). *From Frege to Gödel: A Sourcebook in Mathematical Logic, 1879–1931*. Harvard University Press.

Hilbert, D. (1925). On the infinite. Lecture given in Münster, 4 June 1925. English translation in *From Frege to Gödel* (pp. 367–392).

Houser, N., & Kloesel, C. (Eds.) (1992). *The Essential Peirce. Selected Philosophical Writings Volume 1 (1867–1893)*. Indiana University Press.

Isaac, A. M. C., Szymanik, J., & Verbrugge, R. (2014). Logic and Complexity in Cognitive Science. In *Johan van Benthem on logic and information dynamics* (pp. 787–824). Springer International Publishing.

Jeroslow, R. (1975). Experimental logics and Δ_2^0-theories. *Journal of Philosophical Logic*, 4(3), 253–267.

Kaså, M. (2012). Experimental logics, mechanism and knowable consistency. *Theoria*, 78(3), 213–224.

Kaså, M. (2015). A logic for trial and error classifiers. *Journal of Logic, Language and Information*, 24(3), 307–322.

Kaså, M. (2016). Formally modelling convergent dynamic meaning: Results on compactness and axiomatizability.

Kaså, M. (2017). Analytic tableaux for trial-and-error reasoning.

Lindström, P. (1997). *Aspects of Incompleteness*. Lecture notes in logic (Vol. 10). Springer-Verlag, 1997.

Lindström, P. (2006). Remarks on Penrose's "new argument". *Journal of Philosophical Logic*, 35, 231–237.

Lucas, J. R. (1961). Minds, machines and Gödel. *Philosophy*, 36, 112–137.

Ludlow, P. (2014). *Living Words*. Oxford University Press.

Magari, R. (1974). Su certe teorie non enumerabili. *Annali di Matematica Pura ed Applicata IV*, XCVII, 119–152.

McCarthy, T., & Shapiro, S. (1987). Turing projectability. *Notre Dame Journal of Formal Logic*, 28(4), 520–535.

Mostowski, M. (2008). Limiting Recursion, FM-representability, and Hypercomputations. In *Logic and Theory of Algorithms* (pp. 332–343). University of Athens.

Peirce, C. S. (1877). The fixation of belief. *Popular Science Monthly*, 12, 1–15. Reprinted in *The Essential Peirce. Selected Philosophical Writings Volume 1 (1867–1893)* (pp. 109–123).

Peirce, C. S. (1878). How to make our ideas clear. *Popular Science Monthly*, 12, 286–302. Reprinted in *The Essential Peirce. Selected Philosophical Writings Volume 1 (1867–1893)* (pp. 124–141).

Penrose, R. (1994). *Shadows of the Mind*. Oxford University Press.

Putnam, H. (1965). Trial and error predicates and the solution to a problem of Mostowski. *The Journal of Symbolic Logic*, 30(1), 49–57.

Putnam, H. (1975). The meaning of 'meaning'. *Minnesota Studies in the Philosophy of Science*, VII, 131–193.

Quine, W. (1948). On what there is. *The Review of Metaphysics*, 2(1), 21–38.

Ramsey, F. P. (1927). Facts and Propositions. Reprinted in *Foundations* (pp. 40–57).

Ramsey. F. P. (1978). *Foundations. Essays in Philosophy, Logic and Economics*. D. H. Mellor (Ed.). Routledge & Kegan Paul.

Sahlin, N.-E. (1990). *The Philosophy of F. P. Ramsey*. Cambridge University Press.

Sahlin, N.-E., & Kaså Palmé, M. (2005). Ramsey sentences: an observation. *Metaphysica*, Special Issue 3, 109–117.

Shapiro, S. (1998). Incompleteness, mechanism, and optimism. *Bulletin of Symbolic Logic*, 4(3), 273–302.

Smullyan, R. M. (1995). *First-order logic*. Dover Publications.

Stenning, K., & van Lambalgen, M. (2012). *Human Reasoning and Cognitive Science*. MIT Press.

Waismann, F. (1945). Verifiability. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 19, 119–150.