# A Defence of Wittgenstein's Radical Conventionalism

Ásgeir Berg Matthíasson

University of
St Andrews

This thesis is submitted in partial fulfilment for the degree of

Doctor of Philosophy (PhD)

at the University of St Andrews

July 2020

## Candidate's declaration

I, Ásgeir Berg Matthíasson, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 80,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree.

I was admitted as a research student at the University of St Andrews in September 2016.

I received funding from an organisation or institution and have acknowledged the funder(s) in the full text of my thesis.

Date July 22, 2020 Signature of candidate

## Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

Date July 22, 2020 Signature of supervisor

Date July 22, 2020 Signature of supervisor

## Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Ásgeir Berg Matthíasson, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

**Printed copy**

No embargo on print copy.

**Electronic copy**

No embargo on electronic copy.

Date 22 July, 2020 Signature of candidate

Date 22 July, 2020 Signature of supervisor

Date 22 July, 2020 Signature of supervisor

**Underpinning Research Data or Digital Outputs**

**Candidate's declaration**

I, Ásgeir Berg Matthíasson, hereby certify that no requirements to deposit original research data or digital outputs apply to this thesis and that, where appropriate, secondary data used have been referenced in the full text of my thesis.

Date 22 July, 2020 Signature of candidate

**Abstract**

The first part of this thesis develops a game-theoretic solution to the rule-following paradox, based on Wittgenstein's suggestion in the *Philosophical Investigations* that to follow a rule is a practice. I introduce the notion of a *basic constitutive practice* which I argue can account for the correctness conditions of rule-following and meaning, for indefinitely many cases and without circularity, by identifying correctness with a point on a correlated equilibrium of such a practice. The solution crucially relies on, and makes precise, the Wittgensteinian concepts of training, agreement in judgement and our form of life.

In the second part of the thesis, this solution to the paradox is applied to the problem of mathematical truth. I argue that the essence of Dummett's reading of Wittgenstein as a radical conventionalist is not Dummett's emphasis on decision, but rather the contrast with more moderate forms of conventionalism, whereby an unreduced notion of consequence is appealed to in order to move from stipulated truths to further, more remote truths. Instead of this picture, the radical conventionalist view holds that each truth is a direct expression of the convention and that there is no external criterion at all for the correctness of each step in a mathematical proof except our own practice.

By then identifying mathematical correctness with correctness in basic constitutive practices the view is able to avoid the problems that dogged moderate forms of conventionalism, e.g. Quine's regress problem, as the game-theoretic structure of such practices is able to define correctness for indefinitely many cases without appealing to anything outside itself. Finally, I argue that thus put, common and forceful arguments against radical conventionalism can be answered and conclude that it remains a viable view in the philosophy of mathematics.

*Turing doesn't object to anything I say. He agrees with every word. He objects to the idea he thinks underlies it. He thinks we're undermining mathematics, introducing Bolshevism into mathematics. But not at all.*

—L. Wittgenstein,
to his students in Cambridge,
1939.

# Contents

# Acknowledgements

First and foremost I would like to express my gratitude to my two supervisors, Peter Sullivan and Crispin Wright. I feel enormously privileged to have spent these last few years under their supervision, and to have benefited from their breadth of knowledge and unparalleled philosophical acumen. Peter's generosity with his time and advice has gone well beyond what a graduate student can reasonably expect of their supervisor and his endless patience in the face of my stubbornness has saved me from many embarrassing blunders in the present essay. Crispin's wise counsel, constant encouragement and comments on various drafts have likewise been invaluable. I cannot thank them enough.

I would also like to thank all the other Early Stage Researchers who took part in the Diaphora project, Moritz Baron, Jonathan Egeland Harouny, Michel Croce, Jonathan Dittrich, Matthew Jope, Vincent Grandjean-Perrenoud-Contesse, Slawa Loev, Giada Margiotto, Matheus Valente Leite, Ali Abasnezhad, Lisa Vogt, Tricia Magalotti and Madeleine Hyde for their friendship and camaraderie during this time, as well as Sven Rosenkranz and Chiara Panizza for all their hard work in running the project. Without them, my philosophical life would have been that much harder, if not impossible. I would also like offer my thanks to Anandi Hattiangadi, Åsa Wikforss and Kathrin Glüer-Pagin in Stockholm, and Hannes Leitgeb in Munich for their support and warm welcome during my research visits, as well as for their helpful comments on my work.

All the wonderful people at Stirling likewise deserve my gratitude, including, but not exclusively, Peter Milne, Alisa Mandrigin, Colin Johnston, Michael Wheeler, Alan Millar, Kent Hurtig, Sonia Roca-Royes and Giovanni Merlo, as well as my office mates Paul Conlan, Jimena Clavel Vazquez, Indrek Lobus and Xintong Wei, and, last but not least, Giacomo Melis for his great humour and

companionship. Virginia Respinger, Claire Exley and Zoë Mawby also deserve special thanks for all their invaluable help. Likewise I would like to thank Patrick Greenough, Kevin Scharp and Aaron Cotnoir at St Andrews for their special advice and encouragement, as well as Fenner Tanswell for our numerous discussions on philosophical matters.

I cannot thank my friends and family enough for their friendship and support. They are too numerous to list, but my gratitude is no less for that. Finally, thank you, Julia, for all your support, love and encouragement. Without you, I never would have made it across the finish line.

I dedicate this thesis to the memory of my grandfather, Oddgeir Sigurðsson, who passed away while I was writing it. His good humour, kindness and fortitude will always stay with me.

—Ásgeir Berg Matthíasson.

July 21, 2020,
Berlin, Germany.

# Funding

Work on this thesis has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 675415.

# Publication

Material from Chapter 6 is forthcoming in an article due to appear in the *British Journal for the History of Philosophy* under the title "Contradictions and falling bridges: What was Wittgenstein's reply to Turing?" (Matthíasson, forthcoming).

# Abbreviations of cited works by L. Wittgenstein

AWL *Wittgenstein's Lectures: Cambridge, 1932–1935*, ed. A. Ambrose, Chicago University Press, Chicago, 1982.

BB The Blue Book. *The Blue and Brown Books*, Basil Blackwell, Oxford, 1958.

BrB The Brown Book. *The Blue and Brown Books*, Basil Blackwell, Oxford, 1958.

BT *The Big Typescript: TS 213*, eds. and trans. C. Grant Luckhardt and Maximilian A. E. Aue, Blackwell Publishing, Oxford, 2005.

CV *Culture and Value*, ed. by Georg Henrik von Wright, trans. Peter Winch, Wiley-Blackwell, London, 1984.

LFM *Wittgenstein's Lectures on the Foundations of Mathematics: Cambridge, 1939*, ed. C. Diamond, Cornell University Press, Ithaca, 1976.

PG *Philosophical Grammar*, ed. R. Rhees, trans. A. Kenny, Basil Blackwell, Oxford, 1974.

PI *Philosophical Investigations*, trans. G.E.M. Anscombe, 3rd edition, Blackwell Publishers, Oxford, 2001.

PI *Philosophy of Psychology – A Fragment*, trans. G.E.M. Anscombe, 2nd edition, Basil Blackwell, Oxford, 1958.

PR *Philosophical Remarks*, ed. R. Rhees, trans. R. Hargreaves and R. White, Basil Blackwell, Oxford, 1975.

RFM *Remarks on the Foundations of Mathematics*, eds. G. H. von Wright, R. Rhees and G. E. M. Anscombe, trans. G. E. M. Anscombe, 3rd edition, Basil Blackwell, Oxford, 1978.

# Introduction

Ludwig Wittgenstein is widely considered to be one of the most important philosophers of the 20th century and is so regarded because of his influential work on the philosophy of language and mind. Less well-know is perhaps that most of his unpublished work concerns the philosophy of mathematics and that he himself considered his contributions to that subject to be his most important (cf. Monk 1991, p. 466). Indeed, it might even be argued that his contributions to philosophy in general were mostly conceived of as an outgrowth of this work.

Despite a decent amount of exegetical work on Wittgenstein's philosophy of mathematics, however, the literature on this subject has unfortunately not overlapped very much with contemporary debates in the field more generally, with some notable exceptions, and hence Wittgenstein's influence in contemporary philosophy of mathematics is quite minimal, despite his enormous influence in analytic philosophy as a whole. This is not least because of the difference in style and articulation of the major problems in each tradition and the obscure nature of Wittgenstein's work in this area, none of which was prepared for publication by him.

One of the more important exceptions to this general trend is Michael Dummett's reading of Wittgenstein, articulated in a pair of articles, the first of which was published shortly after the posthumous publication of Wittgenstein's *Remarks on the Foundations of Mathematics* (Dummett 1959) and the second after Dummett's retirement (Dummett 1993).

Despite being published decades apart, Dummett's reading of Wittgenstein is remarkably consistent, with only minor refinements being made in the intervening years. According to this reading, Wittgenstein was as a *radical conventionalist* who held that the "logical necessity of any statement is a direct expression of a linguistic

convention" (Dummett 1959, p. 329) and "not just consequences of conventions, but individually conventional".[1] On this view, our mathematical practices determine, in some unspecified way, for each mathematical proposition individually, that it true or false. Dummett couches his reading in the language of 'choice' and 'decision'—for him, Wittgenstein thought that the necessity of a given necessary truth is always due to our having "expressly decided to treat that very statement as unassailable" (ibid., p. 329) and not because it follows from other conventions we have adopted.

Dummett's reading of Wittgenstein's philosophy of mathematics is a profound one, but despite being well-grounded in the text and expounded by a philosopher with deep knowledge of Wittgenstein's philosophy, it has been almost universally rejected, mostly for philosophical reasons, it seems, rather than exegetical ones, and it often feels as if the reasons for rejecting it amount to nothing more than an *argumentum ad lapidem*: radical conventionalism is plainly false, and therefore it simply cannot be that a great philosopher like Wittgenstein ever held it. This thesis is a defence of Dummett's reading of Wittgenstein as a radical conventionalist, both substantially and exegetically.

I do not want to overstate my case here at the outset, however. Most philosophers have taken this notion of decision to be an intrinsic part of what Dummett's conception of radical conventionalism is, and that undoubtedly motivates their rejection of the view. Instead, I will argue that we should understand the view by contrasting it with more moderate forms of conventionalism whereby each truth is indeed just a consequence of a given convention and not determined individually by the convention. These forms of conventionalism have been subject to devastating objections in the literature, and the radical version is meant to overcome these. The defining aspect of radical conventionalism is therefore that there is no criterion of correctness outside of our practice and that each truth is directly determined by them. Hence, I will argue that the emphasis on choice is neither here nor there, even by Dummett's own lights.

Even though the thesis advances a particular reading of Wittgenstein and a defends the philosophical position that this reading contains, its underlying motivation is to articulate a Wittgensteinian approach to the philosophy of mathematics

1. Putnam 1979, p. 424 as cited by Dummett 1993.

that can contribute to the contemporary debate in a way that working philosophers in the philosophy of mathematics can understand and appreciate. The thesis is therefore written for the "philosophical journalists" Wittgenstein so despised, and not the better kind of reader he himself wished for.[2]

This approach has certain methodological implications. First of all, I try to avoid the jargon and idioms of Wittgenstein scholarship as much as I can, as that is both off-putting to analytic philosophers generally, and in my view, often hides the problems being discussed, rather than illuminating them. In the *Lectures*, Wittgenstein laments: "The seed I'm most likely to sow is a certain jargon" (*LFM* XXXI, p. 293) and the scholarship on Wittgenstein's philosophy ever since unfortunately bears him out to a large extent, as philosophers often come close to using Wittgensteinian terminology as thought-stopping clichés, without clearly articulating what the problems are and how they are thereby solved. Or as Steve Gerrard has recently put it,

> It is one thing to say that "words have meaning only in the stream of life," another to cash out the slogan in a way that is helpful in understanding mathematics. It is one thing to say that mathematics is a motley of language-games, it is another to say precisely what sort of language-games they are, and how they differ from empirical ones. (Gerrard 2018, p. 158)

It is my hope that the account of mathematical truth presented here does not suffer from these defects, and that even if the reader disagrees with it, they will nevertheless understand it and have specific reasons for rejecting it.

Second, I have presented my account of mathematical truth as specific claims and theses, and since Wittgenstein famously rejected philosophical theorising, there is no doubt that "many of my formulations and recastings of the argument

2. An early draft of the Preface to the *Philosophical Investigations* reads:

> With repugnance I hand over the book to the public. The hands in which it will fall are mostly not the ones in which I like to imagine it. May it, I wish, soon become entirely forgotten by the philosophical journalists, and thus perhaps remain preserved for a better kind of reader. (*CV*, p. 66e)

The present author unfortunately considers himself to be a philosophical journalist in this sense, and hence among the less desirable readers.

are done in a way Wittgenstein would not himself approve", as Kripke says about his own work, written in a similar spirit (Kripke 1982, p. 5).

I will not be bothered about this here. There is no doubt that there is a strain in Wittgenstein's work which rejects philosophical theorising as useless and incoherent, and recommends philosophical 'therapy' in order to rid the philosopher of the urge to engage in such work. There is, however, also no doubt that Wittgenstein himself engages in such theorising and puts forth positions that can hardly be described otherwise than as philosophical theories. Here, I am in agreement with David Stern:

> However, if we are to do justice to the full range of positions set out in Wittgenstein's writing, we must acknowledge that Wittgenstein was continually moving back and forth between proto-philosophical theorizing and Pyrrhonian criticism of such theories. As a result, one can selectively marshal texts from every stage of his career that show him defending philosophical theories, and one can also construe those texts as attacking such theorizing. (Stern 2018, p. 139)

Since my goal is to articulate a Wittgensteinian approach to the philosophy of mathematics precisely in the form of such theories, I will therefore set therapeutic readings aside, despite their exegetical value.

The situation is very much the same when it comes to Wittgenstein's later philosophy of mathematics. A large part of what Wittgenstein wrote on the subject was published posthumously in 1956 as the *Remarks on the Foundations of Mathematics* and is a selection of remarks taken from many different manuscripts and typescripts, some of which had not been prepared for publication by Wittgenstein at all. The rest exists as lecture notes compiled by his students (the most important of which was published as the *Lectures on the Foundations of Mathematics*) or as remarks in unpublished manuscripts. This is to say nothing of other compilations of remarks which concern the philosophy of mathematics, such as *Philosophical Remarks* or *Philosophical Grammar*. The nature of this body of work is such that it is highly doubtful that there exists a single coherent view which deserves the title 'Wittgenstein's later philosophy of mathematics'. I will therefore only claim that the reading I advance here is but one strand of Wittgenstein's thought on the

subject, and similarly, that the resulting account of mathematical truth defended here is inspired by him, rather than *his* proper.

**The structure of the thesis** The thesis is in two parts. Part I concerns the rule-following paradox and advances a game-theoretic solution to the paradox which is inspired by Wittgenstein's claim that to follow a rule is a practice. The goal is to make that claim more precise and understandable to analytic philosophers. This part of the thesis is intended to form a coherent whole, independent from the second half. Part II then uses the solution to the paradox advanced in the first part to deflect criticism of radical conventionalism.

In the first chapter, I introduce the rule-following paradox as it appears in Wittgenstein's discussion in the *Philosophical Investigations*, as well as what seems to be his own solution to the paradox. I make three main claims: that (i) Wittgenstein was primarily concerned with arguing against a certain intuitive response to the rule-following problem, namely that it is an act of mind, that of meaning the instruction in a certain way, that provides the solution, and (ii) that Wittgenstein has presented us with a real problem that deflationary responses are not adequate in dissolving, and (iii) that he really is proposing a substantial account of meaning by appealing to our practice in following rules, where training, agreement in judgement and a common form of life play a constitutive role.

I argue that Wittgenstein's own text does not seem present a satisfactory solution to the problem he himself poses, and that the claim that to follow a rule is a practice is too obscure as it stands to adequately constitute the correctness and objectivity of rule-following. I also emphasise that Wittgenstein himself does not seem to have rejected such notions, *contra* some commentators.

In Chapter 2, I discuss the problem from the point of view of its most common contemporary version, that of Saul Kripke (Kripke 1982). Kripke develops the problem as a sceptical paradox that aims to show that there is no such thing as meaning, and offers a 'sceptical solution' that is supposed to make room for this intolerable conclusion. I look at some common ways to respond to his argument, e.g. recent a dispositionalist accounts due to Jarred Warren (2018) and community solutions more generally, and argue that they fail, mostly because the cannot account for the possibility of error, on the one hand, and objectivity in rule-following, on

the other. I remain agnostic about the question of whether Kripke's interpretation gets Wittgenstein right or not, as the main aim of this chapter is to get clearer on what an acceptable solution to the paradox outlined in Chapter 1 requires in the context of the contemporary debate. Nevertheless, I argue that the problem is not best seen a proper paradox, but rather as a challenge to provide a theory of meaning on the one hand, and a powerful strategy to rule out accounts that we otherwise might find most intuitive and plausible. This challenge can be posed quite coherently and without paradox.

Chapter 3 concerns the relationship between meaning and rules. A widespread, almost orthodox, conception of language is that it is an activity constituted and guided by rules. By this, it is meant that the meaning of words are determined by the rules that apply to them and that the rule instructs speakers in how to use words. It is then supposed to follow that learning a language is a matter of grasping the rules governing the use of words and subsequently follow them when performing speech acts.

The chapter is in three parts. In the first two parts, I present two problems for the idea that language is governed by rules, the first due to Crispin Wright, and the second to Katrin Glüer and Peter Pagin. In the last part of the chapter, I argue that contrary to the orthodox reading of his later philosophy, Wittgenstein did not conceive of language as rule-governed, and in fact, the rule-following paradox is one of his main arguments against that view. The idea behind this chapter is to set up the dialectic such that it becomes plausible to look elsewhere for the constitution of meaning. The point is not, however, I should emphasise, that we never actually follow rules, even when speaking a language, but rather that rules do not play an explanatory role in accounting for why we have this ability in the first place, nor for why words have meaning. We do not learn language by learning rules, we learn what it is to follow rules by learning a language.

In Chapter 4, the last chapter of Part I, I introduce my solution to the rule-following paradox. The solution is given in terms of what I will call *basic constitutive practices*. The game-theoretic structure of such a practice constitutes what it is to take part in the practice itself by defining the correctness conditions of our most basic concepts as those actions that lie on the correlated equilibrium of that very practice. This structure can define, without circularity, what it is to follow a rule

for indefinitely many cases, and hence account for the infinitary and open-ended character of rules and meaning. The solution crucially relies on Wittgensteinian concepts such as training, agreement and form of life.

The resulting picture of language will preserve the objectivity and correctness conditions of meaning, all the while providing room for the community as a whole to make a mistake. I also argue that by accepting the account presented, we have a strong reason to reject the idea that language is rule-governed.

The first chapter of Part II, Chapter 5, introduces conventionalism, and contrasts what we might call 'orthodox conventionalism' with radical conventionalism. I also introduce Dummett's various definitions of radical conventionalism. There, I will argue that the essence of radical conventionalism is, both in Dummett's early paper, as well as his later, that each true statement in the relevant to domain is so by convention directly, and not derived from some privileged set of stipulated truths via conventionally chosen rules, as orthodox conventionalism has it. I also argue that Dummett's emphasis on 'choice'—that the correctness of any step in following a rule is the result of a choice we make— is not necessary, and indeed just Dummett's way of speaking of convention more generally. We should therefore rather see radical conventionalism through the contrast with its orthodox counterpart whereby each truth is directly determined by the convention without any external criterion or constraint.

In Chapter 6, I will give a close reading of Wittgenstein's discussion in the *Lectures on the Foundations of Mathematics* (henceforth the *Lectures* or *LFM*), and argue that the position Wittgenstein outlines there, one where true mathematical statements are truths of language and our agreement about the particular case is constitutive of our concepts, meets the definition of radical conventionalism outlined in the previous chapter, and in the final chapter, Chapter 7, I pull it all together and argue that by using the solution to the rule-following paradox developed in Chapter 4, we can give an account of radical conventionalism which is able to withstand the most common and difficult objections to it, including Dummett's influential arguments, and that it is in fact a viable view in the philosophy of mathematics.

# Part I

# The rule-following paradox

# Chapter 1

# Introduction to the rule-following paradox

The notion of a simple function defining a sequence is a familiar one. For instance, we say that the function $f(n) = 2n$ defines a unique sequence—that of the even numbers—and if asked to continue it beyond a finite segment such as '$0, 2, 4, 6 \ldots$' we would be in no doubt as to what further numbers to write down—to an arbitrary length. Of course, we could never write down the whole sequence—that would require writing down infinitely many digits in our finite lifespan—but we feel secure in claiming that a finite expression of the rule has determined, whether it is in the form of such a short segment, written down in function notation or merely as a verbal explanation, for infinitely many places, what we should write down in each case. All this we have learned from finite examples that we extend or project into new and new cases. Indeed, we say that if someone has grasped the rule above and wishes to be in accord with it, she must write down 12 after 10, 242 after 240 and 149 096 after 149 094, and so on, even without us ever having thought about these particular cases in expressing the rule itself. Yet this compulsion is not a physical compulsion nor a psychological one—nobody's hand is forced to write down 12 after 10 nor is anyone unable to write down 13, if they so fancy. It seems—on the contrary—to be a kind of *logical* or *normative* compulsion.

In the so-called rule following considerations, which play a central role in

the *Philosophical Investigations* and touch up on most aspects of Wittgenstein's philosophy—although no consensus exists as to how to read them—Wittgenstein examines what this compulsion amounts to (or on some readings, what it cannot amount to) and what it means to say that a rule determines infinitely many cases. These considerations also play a central role in his philosophy of mathematics, especially regarding the notions of correct inference and proof, and even the ontological status of mathematical objects, although it is even more controversial in this case what work exactly Wittgenstein intends for them to do.

In this chapter, I have two aims: to (i) introduce the rule-following paradox as it appears in Wittgenstein's discussion in the *Philosophical Investigations* and the elements of his own solution, and to (ii) point out certain problems with what that solution seems to be—i.e. argue that Wittgenstein's own text does not, at least on the surface, present a satisfactory solution to the problem he himself poses. This includes the well-know claims that meaning is use and that to follow a rule is a practice.

## 1.1 Background to the debate: meaning mentalism

In his remarks on rule-following in the *Philosophical Investigations*, Wittgenstein seems to be arguing against an intuitive conception of what a rule is and what counts as following one. It is roughly characterised by the combination of the following two assumptions:[1]

(C) At each step in the application of a rule, there are certain actions that are correct and all others are incorrect, independently of any judgement or belief any particular agent might have about it.

(M) In giving or understanding a rule (and thus being able to follow it) an act or state of mind is both necessary and sufficient.

I will call the position characterised by these assumptions *meaning mentalism*.

Meaning mentalism assumes both that what a rule is or requires is independent of the linguistic practices or opinions of rule-followers, and that for someone to

1. (C) for 'Correctness' and (M) for 'Mentalism'.

mean something by an utterance referring to a rule-governed concept, they only have to be in a certain mental state. That is to say, it holds that any statement of the form '*A* means the rule *x* by the symbol *p*' is true if and only if *A* is in some appropriate mental state. In particular, if *A* is developing a series of numbers according to some rule, *A* means this rule, and not another, if *A* is in the mental state associated with this rule or performed an act of mind similarly associated. It follows that at each step, the correctness conditions of *A*'s actions are determined by this mental state: it is correct to write *x* at step *n*, and not *y*, because *x* is in accordance with *that* rule which was meant and *y* not.

There are a number of natural assumptions about meaning a meaning mentalist might make. Here I will only mention a few.[2] First of all, a meaning mentalist would naturally believe that following a rule is simply a private affair. We can explain why *S* is following one rule rather than another by merely referring to his mental states (and thus a fact about *S*: that he is in this particular mental state). This is private because there is no need to refer to other people or linguistic practices in general to explain how *S* follows the rule. It is simply by being in this mental state and nothing else.

It is likewise natural for a meaning determinist to believe that coming to *understand* a rule (or 'grasping it') is a private matter: while my parents or teachers might teach me how to engage in various rule-following practices, they act "only as heuristic aids to an achievement" (Kripke 1982, p. 80) which depends only on me getting into the right mental state, and thus, absent from any outside influence, I *could* have grasped the rule, however unlikely that seems in practice. The existence of other rule-followers is therefore merely a contingent matter when it comes to learning how to follow a rule and not strictly speaking necessary, since anything *might* put me in the right mental state, say a pill or a blow to the head, if I'm lucky.

The philosophical interest of meaning mentalism about rules comes from a further, often tacit assumption, that meaning in general is a matter of following rules, e.g. that for speaker to mean something by an utterance, the speaker must be following the rules that govern the terms that constitute his utterance. This assumption is, most likely, justified by the further assumption that whenever there

2. I have adapted the term 'meaning mentalism' from Kusch's meaning determinism. For a more detailed discussion, see Kusch 2006.

are correctness conditions, there are rules.[3] This gives rise to the following principle regarding speaker meaning:

(M′) *S* means *x* by the symbol *p* if and only if *S* is in some appropriate mental state $\varphi$.

Given the two assumptions, the argument for this is quite simple: if following a rule is a question of being in a particular mental state and meaning is a question of following rules, it follows that meaning is a question of being in a particular mental state.[4] Here, I will assume that meaning mentalism is committed to (M′), with or without accepting the soundness of this argument, or something like it, and thus to the truth of (M′) as well. This explains why philosophers often move effortlessly between discussing linguistic meaning on the one hand and rule-following on the other. I will do likewise for now, unless the context demands otherwise. In Chapter 3, I will argue that this assumption is actually false, and language is in fact not rule-governed.

Of course, it is possible to accept (M′) without bringing rules into the picture at all, and just hold that to mean something it is necessary and sufficient for *S* to be in some mental state. If we assume a principle analogous to (C) for that notion of meaning, however, i.e. assume that meaning nevertheless has correctness conditions, most of the arguments intended to show that there is something problematic about rules could still be made. I will discuss this further in Chapter 2 and Chapter 3.

**The gap between sign and meaning** Now, according to (C), the correctness of a particular outcome or application of a rule, given the rule, is independent of any particular judgement of any particular person or group. Following a rule is thus an objective matter, where some things are correct (in accordance with the rule) and some things are wrong (not in accordance with the rule) and nobody's opinion makes any difference as to which is the case. This has nothing to do with vagueness. There are presumably cases where it is unclear what the right step is or maybe no fact of the matter, but we assume that for the rules we are considering, this is not

3. See Blackburn 1984b, p. 281 for a similar sentiment. I'll discuss this further in Chapter 3.

4. Kripke's reading of Wittgenstein, which I will discuss in Chapter 2, extends this principle to cover any fact about *S*.

the case. The philosophical problems remain, even if we assume that there is no vagueness whatsoever and each step in following the rule perfectly precise (as is presumably the case for arithmetical examples).

Assumption (M), on the other hand, could be said to be based on the intuitive assumption that the expression of a rule, be it written or given in speech, is in itself meaningless and that something is required to give it life: that there is a gap between the expression of the rule and the rule itself. It is thus not necessary that the combination of signs '$f(n) = n^2$' picks out the function $f(n) = n^2$, for instance, but a contingent matter.[5] Likewise, the instruction 'go left at the crossroads' might for instance have been an instruction to go right at the crossroads, if the meaning of the words 'left' and 'right' were swapped. The written or spoken word, we might say, is just a bunch of meaningless sounds or scribbles until they are given meaning by human beings, and consequently, it could have been the case that '$f(n) = n^2$' in fact referred to some other function.[6]

For the meaning mentalist, it is an act of mind which bridges this gap and brings the signs to life. In the case of the speaker, it is an act of meaning that does the job: the formula '$f(n) = n^2$' gets its meaning, that of referring to that particular rule that it does, from a specific act of mind, that of the speaker meaning the rule in the right way. In the case of the hearer, the rule is likewise understood by being in a particular mental state of understanding by which the hearer grasps the semantic content of the rule and internalises it in his mind. This mental state is, we can say, 'occurrent' and has a beginning, duration and comes to an end. It is by being in this mental state that the hearer is able to follow the rule correctly.

It is implicit in meaning mentalism that once a rule is laid down and meant in a particular way, it is already fixed what counts as being in accord with it and

5. See e.g. McDowell 1984; Finkelstein 2000 for a discussion, as well as the following remark by Wittgenstein (*PI*, §431):

> "There is a gap between an order and its execution. It has to be closed by the process of understanding." "Only in the process of understanding does the *order* mean that we are to do THIS. The order — why, that is nothing but sounds, ink-marks.—"

6. Michael Potter gives a good description of this problem in his book on Wittgenstein's *Notes on Logic*: "Symbols are what signs become when we invest them with meaning. When we read the words on the page, we turn them into the living expression of a situation; mere signs, on the other hand, do not yet say anything about the world" (Potter 2008, p. 210).

what not, even before it is ever applied to any particular case and following a rule correctly is first and foremost a matter of grasping the content of the rule: its meaning. The rule is, to use Wittgenstein's metaphor, 'once stamped with a particular meaning' (*PI*, §220), like 'rails invisibly laid to infinity' (*PI*, §219). In particular, when someone is confronted with a new case in applying the rule, e.g. calculating a sum they have never seen before or applying a predicate to a novel object, their competence is to be explained by having grasped the content of a corresponding rule, and if they wish to be in accord with the rule, they have no choice but to react in that one, predetermined way. This picture of 'rails to infinity' is implicit in the combination of (C) and (M) because if (C) is true, the rule has objective *correctness conditions* for infinitely many cases of its application and if (M) is true, it is determined at the time of utterance, through some act of mind, which rule it is that the speaker is in fact following, and hence when that act of mind has taken place everything is already determined for all possible future applications of that particular rule. If it were not, the speaker would simply not be following that very rule.

Since Wittgenstein doubts this picture, many commentators have concluded that his discussion is aimed at showing that (C) is false, and thus that it cannot be said to be an objective matter whether a rule has been followed or not.[7] The arguments for this are varied, and we will look at some of them in due course, but most aim to show both that there is a gap between the expression of a rule and its application and that there are no facts about the speaker that can fill this gap. They conclude that there is therefore no objective matter as to whether or not a rule has been correctly followed or not. I will call this position *meaning scepticism*. Meaning scepticism, as I will use it here, is in effect denying (C), and as I use it, any position accepting (C) is not meaning scepticism.

In these reconstructions of the argument, the notion of 'fact(s) about a speaker' plays the same role as mental states of the speaker plays in my formulation of (M) (and indeed could be seen as a more general form of it) and thus in order for

7. This is undoubtedly because of Kripke's influential reading of Wittgenstein's rule-following considerations. This, I suspect, however, is based on a misreading of Kripke. There is such a character in Kripke's exposition as the meaning sceptic, but Kripke's Wittgenstein does not endorse this position nor does he share the sceptics fundamental assumption of truth-conditional semantics. I discuss this briefly in Chapter 2.

such arguments to work, a similar assumption must be held to be more plausible than (C). In other words, such accounts derive an absurdity from the combination of (C) and a more general version of (M) and conclude that (C) must therefore be false. In the course of the chapter, and throughout the thesis, I will instead argue that Wittgenstein's real target in trying to undermine meaning mentalism, and its accompanying picture of rules as rails, is not (C), but (M), and since it is the combination of the two that gives rise to this image, and (C) is not rejected, Wittgenstein is not concerned with rejecting the objectivity of rule-following.

Meaning scepticism is often conflated with radical conventionalism, the idea that any step in following a rule is directly determined by a convention. That is because on most expositions of it, most notably Dummett's, it is assumed that this entails that we could have *chosen* to go on in any way that we pleased, and (C) is thus put in jeopardy. I will eventually aim to show that one can be a radical conventionalist without rejecting (C)—suitably interpreted, and indeed that this is the most natural reading of Wittgenstein's often cryptic remarks.

I will begin, however, by giving a quick overview of the general shape of how the argument proceeds in in the *Philosophical Investigations*, not as a detailed exegesis, although some cannot be avoided, but to orient the reader before looking more closely at the matter and give an impression of how the rule-following considerations raise a number of issues in philosophy, e.g. about what it is to follow a rule, meaning, the determination of concepts, the giving of reasons, and especially, the challenge these issues pose for objectivity in mathematics.

### 1.1.1 The rule-following considerations: *PI*, §§185–201

The rule-following considerations proper cover roughly §§185–242 of the *Philosophical Investigations*. In the beginning of those remarks, Wittgenstein tells the following story (which has its roots in discussions of 'understanding' and related concepts from §143 onwards): Suppose we have taught someone, through examples and guidance, to write down the series of natural numbers up to 1000. Next, we teach them by the same method to write down series of the form

$$0, n, 2n, 3n, \ldots$$

when they are given an order of the form "+*n*"—that is to say, at the order "+1", they then write down the series of natural numbers. We've given them tests and exercises for these commands up to 1000 and assured ourselves of their competence.

Now we give them the command "+2" and ask them to continue further, beyond 1000. Instead of writing down 1002, 1004, 1006, … as we had expected them to do, they write down the sequence

$$1004, 1008, 1012, \ldots$$

We try to correct them, but they do not understand our corrections. When we tell them that they were supposed to add two, they say that what they did is correct and what they were supposed to, namely adding two.[8] They point at the series and say: "I did it correctly, I went on in the same way". In this case, Wittgenstein says, it would not be of any use to repeat the same way we taught them and he says that perhaps we should say that what comes naturally to them—when they are given this order and hear these kind of explanations—is to do what we do when we are given the order "add two up to 1000, add four up to 2000, add six up to 3000, etc.". What is at stake here, it seems, is that they do not even agree with us what 'the same' is: if one rule is grasped, doing the same is '1002' and if the other rule is grasped, then '1004' is the same.[9]

Wittgenstein's interlocutor reacts with surprise (§186):

> "What you are saying comes to this: a new insight—intuition—is needed at every step to carry out the order '+n' correctly."

Wittgentein's response is to ask how it is decided that one reply is correct, rather than another, to which the interlocutor replies:

8. In Anscombe's translation, the pupil says: "I thought that was how I was *meant* to do it." This translation, even though it is accurate, is slightly misleading in this context because of the connotations of the verb "to mean", which brings to mind the philosophically loaded noun "meaning", which probably isn't supposed to have been brought up by this stage. Wittgenstein himself writes "Ich dacthe, so *soll* ich's machen" where 'sollen' might also be translated as 'should', 'be supposed to' or 'ought to'.

9. Of course, if we assume that we could simply explain verbally what we mean by reference to some other rules, the problem would simply reappear at that lower level—Wittgenstein's assumption, as evidenced by a similar discussion early in the *Investigations*, is that we are discussing such 'basic' rules.

> "The right step is one that accords with the order—as it was *meant*."

Wittgenstein reacts incredulously and asks whether he then also meant that he should write 1002 after 1000, 1868 after 1866, 100 036 after 100 034 and so on for an infinite amount of propositions. The interlocutor responds:

> "No: what I meant was, that he should write the next but one number after *every* that he wrote, and from this all those propositions follow in turn."

Wittgenstein points out that this is essentially just a restatement of the rule in question and does nothing to solve the problem as it was posed, namely the question of what we should take to be in accord with the rule—the criterion of correctness in following the order 'add 2':

> But that is just what is in question: what, at any stage does follow from that sentence. Or again, what, at any stage we are to call "being in accord" with that sentence (and with the meaning you then put into the sentence—whatever that may have consisted in). (*PI*, §186)

Here, Wittgenstein has introduced the notion of *correctness* into the debate. The question is, when given an order such as '+2', what counts as the correct step or, perhaps equivalently, what *should* the person so ordered reply at each step, if they are to follow the order? It cannot be the finite bit of the series that has already been developed which determines this, since any initial segment of a series underdetermines which series it is. For any finite segment of the natural numbers of length $n$, there correspond infinitely many functions, namely the ones that agree with the segment up to $n$ and then diverge. There is nothing in what the teacher showed the pupil, the examples he used, and so on, which is inconsistent with any of these infinitely many functions and what counts as 'going on in the same way' is not decided without some appeal to the present case: if 1002 is correct it is one rule, but if 1004 is, it is another.

There is some function (in fact infinitely many) that corresponds to the pupils aberrant interpretation and there is nothing in what he has seen so far that rules out this as the function that describes the correct way to go on. The question is, how can a finite set of examples represent any particular, determinate rule, if indefinitely

many rules can be extrapolated from such a set? And if that is the case, what settles the right way of going on—which series the pupil is in fact supposed to develop? The same questions can naturally be asked about rules and orders as well, in which case any finite number of examples and training will underdetermine which rule is in play or which order was given, and it seems that if we need to master such basic rules to ever understand such things, we never can.

Wittgenstein's interlocutor tries to solve the problem by referring to the intentions of the teacher—that it is how the order was *meant* that determines what the student should write down at each step. Wittgenstein's reply to this suggestion presents the following dilemma: if I order someone to 'develop the series corresponding to the order +2' and it is supposed to follow from the way I meant the order that writing down 1002 after 1000 counts as having correctly followed the order (and not 1004), then either (a) I 'meant' infinitely many propositions of the form 'after $n$, write $m$' or (b) I meant that for any number $n$, $n+2$ should be written down.

Given the two assumptions of meaning mentalism, neither (a) nor (b) seem satisfactory: if we accept (a) as a proper account of how meaning can be this criterion of correctness, it seems to involve infinitely many distinct acts of mind, namely to mean all of the propositions, 'after 1000, write down 1002', 'after 1002, write down 1004', …, 'after 1866, write down 1868' and so on. This is hardly plausible.

Furthermore, this way of explaining the matter seems to weaken the 'ruleness' of the rule. In what sense is the rule a rule, if it simply stands for each of its instances? That is to say, intuitively we would think that it is an essential part of what a rule such as 'add 2' *is* that 2 comes after 0, 4 after 2, etc., but if we accept (a), it seems like this is merely a contingent matter, that it could very well have been the case that the rule 'add 2' in fact meant 0 after 2, 2 after 4, etc. but 7 following 4. This should be ruled out by the very nature of the order 'add 2' being the order that it is, but on this model, it is a contingent matter that 7 does not follow 4, rather than a necessary truth.

If we accept (b) on the other hand, matters are not much better, since the account given here is nothing more than a restatement of the rule itself, the very thing we are trying to explain, and thus how we meant the order does not seem to be doing any work at all—to say that what determines each step of sequence

defined by the order 'add 2' is that we mean that one should write the next number but one after each number is not an explanation, because that just *is* the rule 'add 2'. We haven't given any explanation where meaning does any work at all. In other words, if we go for option (b), an act of meaning cannot be what determines each step of the series, since the same question can be asked about how the rule is formulated there, leading to a regress.

The dilemma, then, is that if we say that the order 'add 2' corresponds to a particular application because it was meant in that particular way (and further assume that meaning is an act of mind), either we are committed to 'meaning' an infinite number of propositions or we simply meant another rule, which also would need to be meant in a particular way to correctly applied. In either case, we have not given any satisfactory reply.

This story as Wittgenstein tells it, and the challenge it poses—that meaning as an act of mind is not what determines how to apply a rule—seems to presuppose both that we are prone to think that when we mean something, it involves a distinct act of mind by which each application of the rule can then somehow be assessed, and that this is a plausible and intuitive positon in need of correction. These assumptions of Wittgenstein's come out a bit stronger in a parallel discussion of the *Brown Book*:

> I suppose the idea is this: When you gave the rule "Add 1" and meant it, you meant him to write 101 after 100, 199 after 198, 1041 after 1040 and so on. But how did you do all these acts of meaning (I suppose an infinite number of them) when you gave him the rule? Or is this misrepresenting it? And would you say that there was only one act of meaning, from which, however, all these followed in turn? (*BrB*, p. 142)

and a bit later:

> There is a kind of general disease of thinking which always looks for (and finds) what would be called a mental state from which all our acts spring as from a reservoir.[10] (*BrB*, p. 143)

10. Compare the following passage from *Philosophical Grammar*:

In the next two sections of the *PI*, §§187–188, Wittgenstein continues to criticise this conception of rule-following, especially this idea of an act of meaning being necessary for following a rule correctly. There is, however, little indication that Wittgenstein has anything like (C) in mind as his target, as he constantly shifts the attention back onto (M). In §187, his interlocutor points out that he already knew when he gave the order that his student should have written 1002 after 1000. Wittgenstein seems to agree with this, but warns against us being mislead by the grammar of the words 'know' and 'mean'. He writes (§187):

> For you don't want to say that you thought of the step from 1000 to 1002 at that time—and even if you did think of this step, still you didn't think of other ones.

Rather, Wittgenstein says, knowing that this step was the correct one, means that *if* one was asked what number should be written after 1000, then one would have said 1002. He then continues (§187):

> This assumption is rather of the same kind as "If he had fallen into the water then, I should have jumped in after him."

And for this to be true, we require no act of mind to have already have determined it in some way, as it is just a statement about the person in question (their psychological make-up, perhaps) and what they would do under certain circumstances and has nothing to do with correctness of the act.[11]

In what way is the grammar of 'know' and 'mean' liable to mislead us then? Wittgenstein doesn't explicitly say here, but from various remarks (and how the dialectic in the rule-following remarks themselves develops), it seems that Wittgenstein thinks that we confuse the grammar of these verbs with the grammar of the verb 'to think' and for that reason, we think that when we mean something by an utterance, that meaning is completely determined by some act of mind or men-

> The intention seems to interpret, to give the final interpretation; which is not a further sign or picture, but something else, the thing that cannot be further interpreted. But what we have reached is a psychological, not a logical terminus. (*PG*, 145)

11. We can see here a foreshadowing of how Wittgenstein's own account has dispositional elements.

tal state. Wittgenstein returns to this subject by the very end of *PI* in a series of remarks. In §692, he writes, for instance:

> Is it correct for someone to say: "When I gave you this rule, I meant you to …in this case"? Even if he did not think of this case at all as he gave the rule? Of course it is correct. For "to mean it" did not mean: to think of it. (*PI*, §692)

And at §693, Wittgenstein offers what seems to be a summary of his whole discussion of what it is to follow a rule:

> "When I teach someone the formation of the series …I surely mean him to write …at the hundredth place."—Quite right; you mean it. And evidently without necessarily even thinking of it. This shews you how different the grammar of the verb "to mean" is from that of "to think". And nothing is more wrong-headed than calling meaning a mental activity! (*PI*, §693)

Slightly earlier, Wittgenstein makes the same point in a more concrete way, in terms of a 'connection'. He discusses the case of thinking of someone or referring to someone and points out that it is not enough to merely say their name, since this might be variously interpreted (most obviously because most names have many bearers). The interlocutor then points out that there must then be some other connection between my utterance of the name and the person I actually referred to. Wittgenstein agrees:

> Certainly such a connexion exists. Only not as you imagine it: namely by means of a mental *mechanism*.

Now, these remarks do not totally exclude that Wittgenstein's target with his discussion is something like assumption (C)—that there are correctness conditions for the application of a rule that determine every case. Given how the dialectic develops, however, he much rather seems to be emphasising yet again that (M) is simply the wrong way to think about meaning, not that there is no such thing as objective meaning, as the denial of (C) seems to entail—otherwise, Wittgenstein would not be constantly insisting that there *is* such a thing for a rule to determine

unseen steps—what he does question is how to understand the notion of 'determination'.

In section §188, Wittgenstein connects his criticism of (M) with the idea of "rules-as-rails", or as he puts it there, with the act of meaning "already having traversed the steps". He writes:

> Here I should first of all like to say: your idea was that that act of meaning the order had in its own way already traversed all those steps: that when you meant it your mind as it were flew ahead and took all the steps before you arrived physically at this or that one.
>
> Thus you were inclined to use such expressions as: "The steps are *really* already taken, even before I take them in writing or orally or in thought." And it seemed as if they were in some unique way predetermined, anticipated—as only the act of meaning can anticipate reality.

This might seem puzzling, and it is little wonder that Wittgenstein has been understood as attempting to undermine the objectivity of rule-following by these remarks. A couple of questions immediately present themselves, however. First of all, (1) are we really tempted to use expressions such as these, that "the steps of a rule have already been taken", before we take them? Should (2) we understand Wittgenstein's remark to mean that the steps of the rule are not predetermined? If so, isn't that a rejection of (C)?

As a preliminary answer to the first question (1), the answer seems to be both affirmative and negative. It it certainly seems wrong that philosophers in general are tempted to use such expressions as a simple empirical matter, and perhaps the remark is best explained by Wittgenstein being in conversation with himself—that it is *he* who is tempted to use such locutions. First of all, however, this seems to be a natural picture if we accept both (C) and (M)—as explained above. According to (C), there is such a thing as correctness in the application of a rule, independent of any judgement of particular agents, and according to (M), agents follow a rule by grasping and internalising its content. If we combine these two assumptions, we get that this act of mind (or mental state) has already determined each case in virtue of having grasped the content of the rule. It is therefore the act of mind, giving the rule its meaning, which is determining the infinitely many cases—and that seems

to entail that as soon as this act of mind, a temporal process, has taken place, an infinity of steps is thereby determined. Describing this picture as entailing that 'the mind flew ahead and took the steps' does therefore seem quite apt.

Secondly, to say that 'the steps are already taken' seems like a natural thing for a mathematical platonist to say. For such a philosopher, the mathematical reality which the statements of mathematics describe already exists, independently of human practices, and since the ground for the truth a mathematical statement is the mathematical reality, it follows that even before we grasp a mathematical concept, whatever it describes is already settled by the mathematical reality itself. In this way, the mathematical platonist is implicitly committed to something even stronger than meaning mentalism, but is lead to it by a need to explain how we can ever access or 'hook up with' the mathematical reality.[12] And indeed, as we will see later, Wittgenstein does criticise Frege on this point (see Chapter 6).

The answers to the second question—(2)—is a bit harder to come by. I will examine what Wittgenstein means by 'the steps being determined' in more detail in the next section, and throughout the chapter I will argue that, in fact, Wittgenstein's own position was not that the steps of a formula are undetermined in any important sense. For our current purposes, however, it is enough to note that this view, that an act of meaning is gives a rule its correctness conditions, is connected to Wittgenstein's opposition to the idea that understanding and meaning are mental states, a topic he returns to often in the *Investigations*, including the remarks leading up to the rule-following considerations.

In §198, Wittgenstein connects the rule-following of §185 problem with the regress problem of §141. There, Wittgenstein had considered the idea that to grasp the meaning of a word, e.g. the word 'cube', what matters is to have a particular picture in mind, in this case, a picture of a cube. Wittgenstein points out that one can use this image of a cube in various ways and not only apply it to cubes—by having a different, as he calls it, 'method of projection' (*PI* §141). He then points out that if we suppose that we have the correct method of projection before our minds, in addition to the correct picture, it too can be used in various ways, leading to a never-ending regress. Wittgenstein does not explicitly say what the connection is between the idea of a picture before the mind associated with the object it depicts

12. See Kripke 1982, p. 54 on this point.

and the mental model of rule-following, characterised by assumption (C), but presumably it has something to do with the proposed gap between the expression of a rule and the rule itself.

If the meaning of the rule is determined by an act of meaning by the speaker, the hearer cannot directly grasp it, but must only get to it indirectly, so to speak, by interpreting the words of the speaker. And if that's the case, then whatever the hearer has in mind must be similarly interpreted as it can likewise be applied in different ways, leading to a regress analogous to that of §141.[13] In the case of the interlocutor's reply in §186, his explanation of the rule "add 2" as "write the next but one number after" would be an example of such an interpretation: If there's a problem of what "add 2" means, then there's an analogus problem for that explanation, leading to the regress.

The problem of §198 then is that a rule does not seem to give us any direction, since it needs to be interpreted, and this cannot be done without regress. To follow a rule, only one course of action is supposed to be open to us, by assumption (C), but it seems that our picture of rule-following leaves infinitely many ways open to us where nothing in the teaching of the rule makes one alternative more plausible than another—the content of the rule has vanished. We are like Buridan's ass, standing paralysed—equidistant from each possible interpretation of the rule, receiving no guidance from it. This iteration of the problems seems to be *epistemological*: how can we know what rule our student is following? How can a finite set of examples tell us how to go on?

There is a more general problem lurking, however, as Wittgenstein notes later in §198 and in §201, especially.[14] In presenting this modified version of the argument, Wittgenstein approaches the problem from the other direction, namely that not only are there infinitely many interpretations that are compatible with any expression of the rule, but also that any rule can be so interpreted so that any future action can be said to be in accordance with it (even if we were to assume that the former problem were solved). Wittgenstein writes:

> This was our paradox: no course of action could be determined by a

13. See *PI*, §210 for this as a plausible reading.

14. For another philosopher who separates these two aspects of the argument, see Williams 1999, Chapter 6.

> rule, because any course of action can be made out to accord with the rule. The answer was: if *any* action can be made out to accord with the rule, then it can also be made out to conflict with it. And so there would be neither accord nor conflict here. (*PI*, §201)

On the former iteration of the problem, the question was how we could ever know that our pupil has understood the rule we are trying to teach, and thus how the expression of the rule can give any guidance as to what to do, but here, it is a question of how the rule itself could ever determine correct or incorrect action.

There seems to be a question of what exactly makes it the case that a student has carried out our order correctly, since there is always some interpretation of our order which brings his practice, no matter what it is, into accord with the rule—and this generalises, as we will see, to other facts than just mental states. There seems to be no fact at all that can do this job, be it facts about mental states or anything else.

This, as Warren Goldfarb points out (Goldfarb 2012), easily leads to yet a further question. If we cannot say whether or not our student goes on in the right way, how can we be sure *we* are going on in the right way? After all, our own behaviour in the past is finite and we can only justify the rule to ourselves in a similarly finite way. And if there is no way to tell which way is the right way, what makes it the case that there *is* such a thing as the right way at all? If anything we do can be said to be in accordance with the rule on some interpretation, there cannot even be such a thing as 'accord with the rule' (as Wittgenstein notes as the paradox of *PI*, §201) and in this case, there does not seem to be any objectivity left in our rule-following practices in particular. And indeed, at *PI* §186 Wittgenstein asks exactly this question: "How is it decided what is the right step to take at any particular point?". And the answer so far seems to be, if we accept assumptions (C) and (M): It isn't.

This iteration of the problem seems not to be epistemological at all, but to concern the very *constitution* of meaning: 'the possibility of meaning, not our knowledge of it', as Boghossian puts it (Boghossian 1989, p. 515): if we are following a particular rule or engaging in practice governed by a rule, there are certain actions we *should* take and others we should not *in order* to count as being in accord with that particular rule, or as I will claim is equivalent enough for our purposes, that

there is a notion of correctness in play when we follow rules—i.e. the acceptance of (C). Any explanation of rule-following must be able to account for this feature—explain what constitutes this correctness, and so far, meaning mentalism does not seem to be able to cope.

### 1.1.2 Training, practice and agreement: *PI*, §§201–242

The rule-following paradox of *PI* as outlined in the last section was that no matter how we teach someone to follow a particular rule, our examples and instructions will always underdetermine which rule they are meant to follow, and thus there will be some aberrant interpretation of our teaching which fits with their behaviour, no matter what it is. This led to the further thought that since there are infinitely many different interpretations that can be brought into accord with the rule, there seemed to be no fact of the matter at all as to what makes it the case that some actions accord with the rule and some not—and if so, there could be no such thing as following a rule correctly or incorrectly.

In this section, I will give an overview of Wittgenstein's positive remarks about rule-following in *PI*—remarks that revolve around training, practice and agreement in action. In Chapter 6, I will give a similar overview of his discussion in the *Lectures on the Foundations of Mathematics*.

Wittgenstein's interlocutor (as many commentators have as well) gets the impression that Wittgenstein is rejecting (C). In §189, he had already asked: "But *are* the steps then *not* determined by the algebraic formula?". Wittgenstein's reply, while not explicitly denying the determination of each step by the formula, is not an affirmation either, as he simply rejects the question: "The question contains a mistake" (*PI*, §189), he writes. He then goes on to claim that there two ways a formula can be said to 'determine the steps' of a series: as (a) a statement about how people are all trained and educated so they react to the formula '$x^2$' in the same way. For them, he claims, we might say: "the formula determines every step"; and (b) as a statement about the mathematical form of the formula, i.e. to contrast formulas like $y = x^2$ and $y = x^z$—the latter of which determines infinitely many series, each depending on the value of $z$.

Now, it might seem that Wittgenstein is deliberately changing the subject or

overlooking the obvious way the question was meant: we don't want to know anything about child psychology or how some hypothetical group of people was educated, nor the *mathematical* question of whether $y = x^2$ determines only one answer for each $x$ or not (i.e. whether it is a function), but (c) what makes it the case that the combination of signs '$x^2$' refers to to the function $y = x^2$, rather than some other function, and hence how it is decided for every $x \in \mathbb{R}$, what $x^2$ is? If it is not by some act or state of mind, as the combination of assumption (C) and (M) would have it, then what?[15]

Wittgenstein's answer in the next remark (§190) is to consider the intuitive answer that the way the symbol was *meant* determines the correctness of each step in the application of the rule (and remember that in §187, he had not rejected this as a wrong way of speaking, only as misleading if we take 'to mean' to be a particular mental process). Surprisingly, perhaps, he does not seem to be hostile to this idea. He writes (*PI*, §190):

> It may now be said: "The way the formula is meant determines which steps are to be taken". What is the criterion for the way the formula is meant? It is, for example, the kind of way we always use it, the way we are taught to use it.
>
> We say, for instance, to someone who uses a sign unknown to us: "If by '$x!2$' you mean $x^2$, then you get *this* value for $y$, if you mean $2x$, *that* one."—Now ask yourself: how does one *mean* the one thing or the other by "$x!2$"?
>
> *That* will be how meaning it can determine the steps in advance.

Despite appearances, there seems to be then, on Wittgenstein's view, a way for a formula to determine the steps of a series (that the indexical 'that' refers to nothing is hardly persuasive) which has something to do with how it is 'meant'. The preceding discussion, however, shows that his position is that meaning does after all determine the correctness conditions of the rule, but that it cannot be reduced to

15. If we are tempted to say that it is because $x$ takes all and only the real numbers as values, Wittgenstein can always ask what makes it the case that we are referring to the real numbers when defining the domain of the function? Presumably, we also learnt them by seeing a finite series, and so the problem comes back with regard to the symbol '$\mathbb{R}$'.

mental states. He hasn't, however, told us anything constructive about meaning, except a vague reference to training and 'what we always do', and characteristically, leaves it to the reader to answer his crucial rhetorical question. In Chapter 4, I offer an account of meaning that can be seen as an attempt to answer this question.

After some (puzzling, I admit) remarks about machines, their action and how they can be seen to symbolise their own action, Wittgenstein returns to the paradox. The interlocutor asks: "Then can whatever I do be brought into accord with the rule?" (*PI*, §198) and Wittgenstein replies:

> —Let me ask this: what has the expression of a rule—say a sign-post—got to do with my actions? What sort of connexion is there?—Well, perhaps this one: I have been trained to react to this sign in this particular way, and now I do so react to it. (*PI*, §198

But this is not enough:

> But that is only to give a causal connexion; to tell how it has come about we now go by the sign-post; not what this going-by-the-sign really consists in. On the contrary; I have further indicated that a person goes by a sign-post only in so far as there exists a regular use of sign-posts, a custom. (*PI*, §198)

In the next remark (§199), Wittgenstein considers the question of whether it is possible for only one person to follow a rule only once. Wittgenstein's answer is in the negative—this is not possible, because:

> —To obey a rule, to make a report, to give an order, to play a game of chess, are *customs* (uses, institutions).
>
> To understand a sentence means to understand a language. To understand a language means to be a master of a technique. (*PI*, §199)

As we saw above, Wittgenstein restates the paradox in §201—that no action can be determined by a rule, because any action can be made to fit with the rule expressed so far. Immediately afterwards, however, Wittgenstein seems to reject it as a "misunderstanding". He goes on:

> That there is a misunderstanding here is shown by the mere fact that in this chain of reasoning we place one interpretation behind another, as if each one contented us at least for a moment, until we thought of yet another lying behind it. For what we thereby show is that there is a way of grasping a rule which is not an interpretation, but which, from case to case of application, is exhibited in what we call "following the rule" and "going against it" (*PI*, §201).

In §202, Wittgenstein concludes:

> And hence also 'obeying a rule' is a practice. And to *think* one is obeying a rule is not to obey a rule. Hence it is not possible to obey a rule 'privately': otherwise thinking one was obeying a rule would be the same thing as obeying it. (*PI*, §202)

It is quite common among interpreters to conclude from these remarks that Wittgenstein's concern in his remarks on rule-following is to undermine "the misconception that rule-following is always grounded in (or implicitly contains) acts of interpretation", as Fogelin puts it (Fogelin 2009, p. 22). That's undoubtedly true, but Wittgenstein also seems to be offering his own, positive account in these remarks, as well as in subsequent ones. To follow a rule, he seems to be saying, is a *practice*, something we do, a pattern of behaviour that is exhibited in what we *call* following the rule.[16]

This leaves a twofold puzzle: First of all, how does the appeal to practice help? In Wittgenstein's own example, the student has never seen the case he's presented with before, and hence we can construct a similar case were no one has ever taken that step in following the rule before, and hence our practice has not settled *that* case. It seems that Wittgenstein is saying that whatever we do will be correct, since *that* will then be our practice—but that reply is both highly counterintuitive and undermines the notion of correctness itself: we wanted something that explained why one action is correct and another incorrect, not merely declare whatever we do as correct.

Second, what does the correctness of the rule have to do with what we *call*

16. *RFM* VI, §29: "Following a rule is a human activity."

correct? Surely correctness is more than just being called correct, even by many or all?[17]

Wittgenstein takes this challenge head on in the last three remarks of what I have delineated here as the rule-following considerations (§§240–242). In §241–242, he writes:

> "So you are saying that human agreement decides what is true and what is false?"—It is what human beings *say* that is true or false; and they agree in the *language* they use. That is not agreement in opinions but in form of life. (*PI*, §240)

> It is not only agreement in definitions, but also (odd as it may sound) agreement in judgements that is required for communication by means of language. This seems to abolish logic, but does not do so. It is one thing to describe methods of measurement, and another to obtain and state results of measurement. But what we call "measuring" is in part determined by a certain constancy in results of measurement. (*PI*, §242)

But *prima facie*, this doesn't really answer our worry. If it is impossible for an individual to follow a rule privately, because then there will not be any such thing as correctness, it is hard to see how bringing other people into the picture helps. Wittgenstein seems to be saying that if we evaluate the rule-follower in the context of a community, however we spell that out, there is such a thing as correctness and incorrectness—this is not due to an agreement in opinions, but in *judgements*, as well as in using certain language and having the same form of life.

But what kind of agreement is that? And doesn't the same problem not just re-appear at the level of the community? Will then not anything that the community does be correct, and isn't there such a thing as the community itself being wrong? And doesn't our practice itself sometimes help itself to the fact that everyone can be wrong (or was in a particular case)? If Wittgenstein is right, it seems that this distinction is meaningless, which it is not.

17. The first element of this reply roughly corresponds to what Dummett has called 'radical conventionalism' and the second to Kripke's sceptical solution. See Chapter 5 and Chapter 7 for discussion on the former and Chapter 2 for the latter.

Crispin Wright has summarised the conundrum Wittgenstein puts us in as follows:

> So we have been told what does not constitute the requirement of a rule in any particular case: it is not constituted by our agreement about the particular case, and it is not constituted autonomously, by a rule-as-rail, (our ability to follow which would arguably be epistemologically unaccountable.) But we have not been told what does constitute it; all we have been told is that there would simply be no such requirements were it not for the phenomenon of actual, widespread human agreement in judgement. How can he possibly have thought that this was enough? (Wright 2001b, p. 168)

And yet, in much of the literature on Wittgenstein's philosophy of mathematics (or indeed his philosophy in general) it is very much assumed that we do have a grip on how our practice, techniques, custom and forms of life can account for the correctness conditions of these same practices in future cases. This aspect of Wittgenstein's thought is, again in the words of Wright, both "familiar, and ill understood" (Wright 2001e, p. 188), but at the same time, these pieces of Wittgensteinian jargon are, as Ian Hacking put it, "often cited as if they were at the end, not in the middle, of a series of thoughts" (Hacking 2014, p. 2). The fact is, that Wittgenstein's account *is* unclear and is far from obvious what all this is supposed to amount to. Appealing to our common practice and forms of life is just not enough.

What, then, about the well-known claim that meaning is use (*PI*, §43)? First of all, we should note that Wittgenstein never makes the claim that meaning is use in the context of rule-following. It is therefore likely that if he did think that the idea that 'meaning is use' was relevant for the rule-following paradox (and there is no compelling reason to think otherwise), then he probably would have thought that the claim that meaning as use and the claim that to follow a rule (or indeed to mean something) is to take part in a practice are somehow equivalent (since otherwise he would have brought it up again), and so there is still a lacuna to be filled in the account.

If meaning and use and meaning as practice are not meant to be equivalent,

however, it would still be unclear how taking meaning to be use will help. It is quite unclear what it means to say that the teacher uses the symbol '+2' on that particular occasion so that it is thereby correct for the student to write down 1002, and incorrect to write down 1004. Presumably, it cannot be the teacher's intention or mental state that makes that distinction, given what we've said above, and so we would be hard-pressed to point to anything about the teacher's use in that case that can make it. *Prima facie*, when the appeal to mental states is removed, we are only left with observable behaviour to make the distinction, and to use the term '+2' so that it means one rule looks the same as using it to mean another.[18]

The challenge for any coherent explication of Wittgenstein's philosophy of mathematics, particularly if the goal is to have something to say in contemporary debates in the philosophy of mathematics, is to provide an explanation of how these notions can do the work they are meant to do (i.e. by not merely paraphrasing Wittgenstein's own cryptic discussion as commentators are wont to do) but at the same time preserve something like (C) without falling prey to the rule-following paradox itself. And since the solution (or dissolution) of the rule-following paradox plays such a large role for Wittgenstein's philosophy of mathematics, it becomes vital to be able to answer the following two questions: (i) how should a minimally successful Wittgensteinian account of the normativity and correctness of rule-following and meaning look like, and (ii) what role should the key concepts he appeals to, 'practice', 'technique', 'custom', 'forms of life' and 'agreement in judgements' play in this account?

In the first part of the dissertation, I will try to clarify what is needed in order to come up with a satisfactory answer, before giving my own solution using these concepts in Chapter 4. The second part will be an application of that account of rule-following to the problem of mathematical truth.

18. I don't mean to say that these are the only two options, but if there is a third, Wittgenstein has not told us what it is. For more extensive criticisms of the idea that meaning is use, and arguments against the idea that Wittgenstein meant his remark as a theory of meaning, see Unnsteinsson 2016.

## 1.2 Deflationary accounts of rule-following

Wright has pointed out that there are "tempting deflationary responses" to the paradox (Wright 2007, p. 482): if there is a question of what constitutes the meaning of given word or the correctness conditions of a rule, we could simply say that if a rule is clearly formulated and general enough, that it determines what to do in indefinitely many cases because that's what a rule *is*. Wright writes:

> There is an understandable tendency of philosophers to miss the issues here. For a rule, it may plausibly be said—at least any rule of sufficient generality and definiteness—is nothing if not something that precisely does mandate (or allow) determinate courses of action in an indefinite range of cases that its practitioners will never have explicitly considered or prepared for. So there cannot be a puzzle about how a rule does that, or what settles what its requirements are. To ask how it is settled what complies with the rule is like asking how it is settled what shape a particular geometrical figure has. Shape is an internal property of the figure. What settles what shape the figure has is simply its being the figure it is. (ibid., p. 482)

In other words: the answer 1004 after 1002 is correct because that's what following the rule +2 is. If anything else were correct, it would be some other rule. Similarly, with meaning: if *S* meant *red* by his utterance 'red', then only certain things count as being red, namely the red ones, because that is what being red is. There is then no problem of explaining meaning at all. This way of responding to the paradox we could also call a *constitutive* response, as the steps of the rule are taken to be constitutive of the rule itself.

This response is somewhat intuitive, at least I cannot see myself giving any other response, if ordered to develop the series given by the order 'add 2' and then asked why I wrote down 6 after 4, than "That is what adding two is!". But Wright points out that this is not so simple:

> Yet the concerns merely reformulate and re-assert themselves. If a (suitably precise and general) rule is—by the very notion of 'rule', as it were—intrinsically such as to carry predeterminate verdicts for an

> open-ended range of occasions, and if grasping a rule is—by definition—an ability to keep track of those verdicts, step by step, then the prime question merely becomes: what makes it possible for there to be such things as rules, so conceived, at all? I can create a geometrical figure by drawing it. But how do I create something which carries predeterminate instructions for an open range of situations that I do not think about in creating it? (Wright 2007, p. 483)

He then adds, "What gives it this rather than that content, when anything I say or do in explaining it will be open to an indefinite variety of conflicting interpretations? How can I make it, rather than a competitor, into an object which I intend to follow?" (ibid., p. 483).

Furthermore, it seems that we can still ask the question Wittgenstein poses in his dialectic: if a student is shown by his teacher a set of examples and then expected to go on in the same way, and some ways of going on are correct and some incorrect, whence the correctness? The deflationary response might do well in explaining that generally, but when it comes down to particular instances of rule-following, it falters—sure, 1004 is correct after 1002 because that is what the rule +2 *is*, but what makes it the case that the teacher meant '+2', rather than "+2 up to *n*, +4 afterwards"—in which case case 1004 would have been correct? That rule is a well-defined one, and the teacher *could* have meant that one, rather than +2. The question then becomes: how is it that *S* can mean one rule rather than another with his words? How does the sign '+2' refer to the command +2? Saying that there is an internal relation between a rule and what counts as following it doesn't seem to answer *that* question.

In this section, I want to take a look at one prominent interpretation of Wittgenstein's remarks that in many ways takes a deflationary, or constitutive, approach: that of Hacker and Baker (Hacker and Baker 1984a, 1984b, 2009). I will not claim that Hacker and Baker get Wittgenstein wrong—on the contrary, I rather suspect that they get him right. Instead, I will argue that, while their reading might be based on sound exegesis, they neither solve the problem of correctness, nor point to a way of *dissolving* it, as they sometimes claim to have done. For convenience, I will refer to their reading of Wittgenstein as the *Oxford interpretation*.

### 1.2.1 The Oxford interpretation

The starting point of the Oxford interpretation can be said to be Wittgenstein's rejection of rule-following as interpretation in §201. As we saw above, the thought that following a rule is always an interpretation, an act of mind to mirror the original act of meaning, is a natural one, given the arguments that motivate the rule-following paradox, in particular the supposed gap between the symbol that represents the rule and rule itself. For Hacker and Baker, however, this reasoning is motivated by the idea that there could possibly be a gap between a rule and the practice of following it, and this assumption is, on their view, confused. They write:

> The appearance of a logical gulf between a rule and its extension arises from the mistaken assumption that understanding a rule is at least partly independent of how it is projected onto actions in practice. But however it is formulated or explained, a rule is understood, other things being equal, only if it is correctly projected, i.e. projected in the manner which, in the practice of following the rule, counts as following it. To be ignorant or mistaken about what acts accord with it is to be ignorant or mistaken about what the rule is. (Hacker and Baker 2009, p. 94)

In other words, the relationship between a rule and its application is according to the Oxford interpretation *internal*—and as such this reply is a kind of deflationary or constitutive one. For Hacker and Backer, a rule is internally related to its application, and from there it follows trivially that a rule can never be separated from the practice of following it, since that is what the practice of following a rule *is*. Therefore, according to the Oxford interpretation, what is rejected in §201 is not the "truism that rules guide action" nor that we are "actually applying expressions in accord with their explanations" ("their rules", they add) but rather the misguided idea that

> a rule determines an action as being in accord with in only in virtue of an interpretation. (Hacker and Baker 1984a, p. 420)

Not only do rules guide our action on their view, but they do it by providing a standard 'against which to evaluate performances as correct or incorrect' (Hacker and Baker 2009, p. 66-67). Likewise, appealing to the mental states of the rule-follower to provide this standard of correctness is for them a confusion Wittgenstein sought to excise (ibid., p. 81). In the terms I've used here, the Oxford interpretation endorses (C) and rejects (M).

In the place of (M), however, nothing else is stipulated to bridge the supposed gap between the expression of the rule and the rule itself. We need no "mediating entities" to explain rule-following (ibid., p. 87). Instead, we can make the relation between a rule and what counts as being in accord with it "perspicuous" by clarifying the grammar of the relevant concepts. For instance, if a rule says "add two", then if I add two, I will have done what the rule requires of me and if I add any other number, I will have broken the rule (or disobeyed the order). In general, they say, it is a grammatical truth that "an *F*'s *V*-ing in circumstances *C* is an act that accords with the rule that *F*'s should *V* in *C*" (ibid., p. 87)—that is to say: if I add two when given the order add two, I've followed the rule add two and if I don't I haven't. The rule and the act that is in accord with it cannot be separated, and this is, as they explain, because like

> the relation between a true proposition and the fact that makes it true, the relation between a rule and an act in accordance with it is internal. *This* rule would not be the rule that it is, nor would *this* act be the act that it is, if this act did not *count* as being in accord with this rule.

They go on to say that it is a "grammatical platitude" that a rule determines what counts as following it and that for this reason, it is a mistake to think that a rule does not determine what is in accord with it. The rule and its extension are internally related and hence, the rule 'Add 2' would be a different rule from the rule that it in fact is, if 1002 did not follow 1000. It is, they say, "in language that the rule and the act that accords with it [...] make contact" (ibid., p. 88) and that the rule and "its application are internally related, for we define what following this rule consists in by reference to this result" (ibid., p. 129).

They grant however that this is not the end of the matter. What remains is to explain how such internal relations are possible and this they do by referring to

our ability to master techniques and take part in various practices involving rules—what we *do* when we follow rules. For Hacker and Baker, language is "a rule-governed symbolism" (ibid., p. 135) and it is through the mastery of techniques that we are able to "grasp a regularity in certain rule-governed practices" (ibid., p. 140). Counting, calculating, measuring, weighing and inferring are all examples of techniques and to be able to take part in practices involving them is "to display the ability to use and follow the various rules that define them" (ibid., p. 140) and this mastery is 'manifest' in "a particular act of applying it, but only against a complex background of behaviour exhibiting his abilities and comprising *a practice*" (ibid., p. 141).

The crucial question is, they say,

> what connection is there between an expression of a rule and one's *action*? And the answer is that one may have been trained to respond to a given rule-formulation in such-and-such a way, that there is a regular use of this formulation of a rule as a standard of correctness and a custom of regularly going by it in this way. (ibid., p. 28)

And a bit later:

> To follow a rule is to engage in an activity that exemplifies a regularity recognized as a uniformity. It is a custom (usage, institution), a normative practice. For a kind of behaviour to constitute a case of following rules, it must be embedded in a diachronic setting of normative activities. (ibid., pp. 28–29)

They conclude by noting that the paradox is generated by not realising that not "all understanding can be interpreting" and that how one understands a rule is "exhibited not only in how one interprets it, but also in what one does, in the behaviour that is called following the rule and going against it" (ibid., p. 29).

They emphasise, however, *contra* most commentators, that this does not mean that the notion of "practice" is essentially a social one—that only a practice which is *essentially* social can provide correctness conditions. Hacker and Baker correctly point out that we do not make use of people's agreement to settle whether a rule has been followed or not (we don't say: "2 + 2 = 4 because everyone says so") nor to we

appeal to statistics about people's beliefs, nor do we *define* "correct" in terms of what is *normal* (Hacker and Baker 2009, p. 150). They further claim that using "what is standard or normal in a community" to define what is correct is incompatible with the idea that a rule is internally related to the acts that accord with it (ibid., p. 150–151).

Instead, it emphasises practice as something which people do through time, rather than together. After restating Wittgenstein's claim that how one understands a rule is exhibited in how one follows it, Hacker and Baker write:

> Hence following a rule is an activity, a *Praxis*. It is a misinterpretation to take 'Praxis' here to signify a social practice. The contrast here is not between an aria and a chorus, but between looking at a score and singing. (Hacker and Baker 1984a, p. 420)

They go on to say that the distinction is the same as in the phrase "in theory and in practice" and that Wittgenstein's point was not that rule necessarily require a community of rule-followers, but rather that "words are deeds". The add, however, that "practice" should not be understood as "mere action", but rather as "a regular action in accord with a rule" (ibid., p. 420).

The account so far is quite problematic, however. First of all, Hacker and Baker have told us that to follow a rule is a practice— to "be embedded in a diachronic setting of normative activities" (Hacker and Baker 2009, p. 29). That is to say, the normative aspects of the rule are in part to be explained by rule-followers taking part in *normative* practices. But they have also told us that such practices are rule-governed, and presumably that means that what is correct or incorrect in such a practice is so because we are guided by rules which serve as the criterion for correct action. (ibid., p. 50) If this is not patently circular, it is hard to see how this either *explains* rule-following or *dissolves* the problem as posed by Wittgenstein.

The answer to this puzzle, I presume, is supposed to lie in the constitutive nature of the Oxford interpretation—that there is an internal relation between the rule and the practice it governs. Hacker and Baker often express this by using locutions such as '*x* is correct because *x* is what we call following *that* rule' or 'only *x* counts as following *this* rule'. This is presumably meant to solve the correctness challenge by appealing to the fact that only one way of continuing the series *is* that

particular series—if we did something else, we would be developing a different series. But this hardly helps: if I have seen a finite set of examples of what accords with the rule, say '0, 2, 4, 6…', how do *I* know what *is* called the series of even integers and not something else? Knowing how to follow a rule on this account seems to presuppose that I already know all the steps! How do I project what I have learnt into new and new cases?

The following passage, regarding rule-following in the context of inference, seems to offer an explanation that can avoid this conclusion:

> We are instructed in the techniques of inferring and compelled by our teachers and peers to adhere to them. A specification of what conclusion to draw from given premises is an intrinsic feature of the technique of inference. Hence we are not forced by logic to draw a conclusion – no matter what conclusion we draw, logic will not seize us by the throat! That such-and-such a conclusion follows (or even: 'follows necessarily') is not a form of compulsion. Rather, we are constrained in our judgements about what it is to be called 'a correct inference'. Wittgenstein's account of inference does not derogate from the inexorability of logic. It merely eliminates misconceptions of it. (ibid., p. 6)

Here, we are told that the reason why we say certain things rather than others ("are constrained in our judgements") is that we are instructed in certain techniques and corrected by others.

This, however, doesn't really help, since this will only tell us what I *will* do, but not why what I did was correct. After all, anything a might do will accord with *some* rule, but how can I know what I was supposed to do to follow *this* one?

Perhaps we can see better what this objection to the Oxford interpretation amounts to by the following: rule-governed practices, mathematical practices most prominently, are *open-ended* in that they apply to an indefinite range of cases. If certain outcomes are constitutive of the rules themselves, our practice still underdetermines *which* rules it is that we are actually following, and thus cannot provide a standard of correctness in how to continue. For instance, the rules of arithmetic stipulate that $2 + 2 = 4$ and while it is not hard to see how this outcome is con-

stitutive of our concept of addition, this becomes problematic when it is applied to higher numbers, since we need some way of moving to new and new cases we don't know what to say about. I don't know if, say, 135 664 + 37 863 = 173 527 is true or not, and can therefore not tell if that is a correct application of the rules either, independently of the calculation I've made to reach the conclusion, itself a rule-guided practice. If it is in fact true, then that is the mark of me following the the rules of arithmetic correctly, according to the Oxford interpretation.

However, according to that same interpretation, if I'd write down something else, I'd be following a different rule, say the rules for quaddition, since that outcome would be constitutive of *those* rules. The question is, given all my addition-like practices and all the training and corrections I have received, all of which are finite, which rule am I actually following? If 135 664 + 37 863 = 173 527 is correct, it is addition, if it is incorrect, it is not addition, but everything I have done so far underdetermines which it is. It is not enough to just *stipulate* that I'm adding, and therefore it is 135 664 + 37 863 = 173 527 which is correct. The Oxford interpretation can, in other words, explain well enough what it is to make a mistake in following a particular rule, but it fails to explain what it is to follow the *correct* rule in a given practice.[19]

Hacker and Baker are aware of this objection, or a version of it. They write:

> The same action cannot be treated now as correct, now as incorrect (in the same circumstances) within a single coherent practice. If what was done previously was correct, then surely doing the same thing again must be correct. And whether the same act is performed twice can be settled by inspection. Conversely, if what someone does is unprecedented in a given rule-governed practice (e.g. adding two unprecedentedly large numbers), then it must apparently fall outside the scope of the technique exhibited in the practice. Hence, the question of its correctness must be either nonsensical or open. And this thought leads immediately to the idea that *new cases* relative to a practice cannot be predetermined by the relevant rule. (Hacker and Baker 2009, p. 146)

19. In Chapter 7 I will respond to a version of this objection relating to my own account.

Now, these two objections are not identical, but they can be quite easily converted into one another. It is implicit in my version of the objection, that in order to do what is an accordance with what we have been doing before in our addition-like practice that we must always have been doing the same thing and to project that practice into new and new cases, we must continue to do the same thing. The question is, what determines what counts as the same thing, relative to a rule-governed practice? The constitutive account cannot give an adequate answer, since *anything* we might do in a new case is constitutive of *some* rule, and thus correct relative to *that* rule, and since the finite practice underdetermines which rule we are following, anything we do will be correct.

Unfortunately, however, Hacker and Baker's own reply to this objection is not adequate and misses the point. The first point they make is that we cannot step outside of the practice to find some external criterion to determine whether or not we did the same in two different cases. It is, they say, the practice itself which "is the arbiter of what counts as doing the same thing" (ibid., p. 146) and when someone learns how to follow a rule, they are learning what accords with that rule at the same time, and thus the rule and the concept 'go on in the same way' relative to the rule are intrinsically linked and cannot be separated. What counts as following the rule within "a normative practice", they conclude, "is determined from the perspective of the practice itself" (ibid., p. 146).

While their point that a rule and what counts as the same in following it cannot be separated is well taken, and can indeed be seen to be one of the main lessons of the rule-following paradox, it is hard to see how this answers the objection.

Consider an open-ended practice such as developing the series '+2', and suppose *A* and *B* have a dispute about whether or not one should write 1002 after 1000 or 1004. Further suppose that this is a genuinely new case. Both ways of continuing fit with *some* practice of developing the series and *some* rule of how to go on. Both the practice and the rule up to that point underdetermine *which* practice it is and can therefore not settle the dispute between *A* and *B*—they are both right, given the practice they take themselves to be engaged in. If we think that it is an objective matter which of them is correct, the Oxford interpretation offers no answer: simply saying that 1002 *is* what following the rule +2 is in that case is not a reply *if* the case in question is genuinely a new case, as the dispute

concerns *which* rule we are in fact following. This does not mean that we should be looking for something external to our practice to explain the phenomenon of rule-following, but pointing out where we cannot find an explanation is not the same as giving one.

Perhaps, however, we should understand the Oxford interpretation in an even stronger way: the internal relation is between the speaker's *knowledge* of the rule and their application of it. In an earlier work, Hacker and Baker seem to have understood matters in that way:

> What these premises of rule-scepticism share is a failure to acknowledge that acting in certain ways (what is called 'acting in conformity with the rule') are criteria for understanding a rule, and that acting otherwise is a criterion for failing to understand it…The rule-sceptic distorts this internal relation between acts and rules by treating acting in accord with a rule as making understanding the rule merely a probable hypothesis. (Hacker and Baker 1984b, p. 103)

However, as Anandi Hattiangadi points out, this requirement is simply too strong. If understanding a rule necessarily means that a speaker acts in accordance with it, there could not be any such thing as wilfully doing something else than the rule requires or simply make a mistake—the student might very well understand the rule +2 and still write that 1004 follows 1002, for myriad of reasons. As she puts it, "it is not invariably the case that my behaviour manifests my understanding" (Hattiangadi 2007, p. 175).

Even if we stick with the weaker claim, however, Hacker and Baker's further counter-arguments likewise miss the point. For instance, they point out that how we use the phrase "going on in the same way" is always ambiguous. We might say that when I added 15 to 27 yesterday, I did something different than I when I did it today, or we might say that I did something different when I added 15 to 27 than when I added 15 to 68, or we might say that I did the same. These uses, they point out, are all perfectly all right and not in conflict with each other: it is simply a confusion to think that there is a a "single, correct, context-free, purpose-independent answer to the question of whether this is doing the same as that" . The concepts of "regularity", "predicability" and "agreement" must be grasped from the

point of view of the practice, since there is no way to separate the rule, the practice of following it and what counts as being correct.

These considerations, they say "bear on puzzles about applying rules to new cases". It is true, they say that in a sense every application of a rule is a new case, that is to say, a different one from previous applications—either because the numbers are different or the day is different, etc. In another sense, every application of the rule is an old case, because one is doing the same as before, even if the numbers are larger than anyone has ever added before. The conclude:

> In maintaining that the rule of addition does not predetermine what counts as the correct result of adding two numbers of unprecedented size, we in effect assert that since adding these numbers is doing something different from any previous addition sum, we cannot say that the rule of addition is applied here (in the same way). But that makes no sense (PI §227).[20]

But it is unclear how this answers the puzzle. Even if we agree that there are confusions here about what "doing the same" means, we do not seem to have made much progress: there is still the question of which rule we are following when doing "the same thing"—if we are in fact following *this* rule, it is that, and if *that*, the other thing. Hacker and Baker's arguments simply have not shown that there is no problem in this regard, nor that it can be easily dissolved by pointing out that there is an internal relation between a rule and what counts as following it.

Invoking Wittgenstein's authority is simply not an answer, even if what they describe does not in fact make sense. Nevertheless, their point that the a rule and what counts as following cannot so easily be separated is persuasive. Ultimately, our solution to the paradox must be sensitive to that point.

20. The remark they cite is as follows (*PI*, §227): "Would it make sense to say: "If he did something different every time, we wouldn't say he was following a rule"? That makes no sense."

# Chapter 2

# Kripke's version of the rule-following paradox

In the previous chapter, I argued that Wittgenstein's target with his discussion of rule-following was a certain intuitive conception of meaning and rule-following I called meaning mentalism. It was constituted by the following two assumptions:

(C) At each step in the application of a rule, there are certain actions that are correct and all others are incorrect, independently of any judgement or belief any particular agent might have about it.

(M) In giving or understanding a rule (and thus being able to follow it) an act or state of mind is both necessary and sufficient.

Saul Kripke's famous interpretation of Wittgenstein's discussion of rule-following (Kripke 1982) can be seen as developing the paradox with (M) replaced with a more general version:[1]

(F) For a speaker *S*, there is some fact about *S* that constitutes *S*'s meaning by a given utterance.

This view of meaning entails a very similar picture of language as meaning mentalism, one where some fact about *S* plays the role of *S*'s mental state in meaning

1. (F) for, naturally, 'Fact'.

mentalism: if *S* meant *p* by his use of a symbol $\varphi$, then there is some *fact* about *S* that makes this the case. Furthermore, it is generally held that this fact is both necessary and sufficient for *S* to mean what he means, so whenever that fact obtains, *S* meant *p*.

This more general picture of language, I will call *meaning factualism* (or *factualism* for short) as an analog with 'meaning mentalism'.[2] Meaning factualism is a stronger view than meaning mentalism and the latter is a kind of fact determinism—for the meaning mentalist, mental states *are* the fact we appeal to. Like its cousin, meaning factualism comes with a number of assumptions that likewise are meant to be intuitive and truistic (and in Kripke's discussion, it is often supposed that the fact that the sceptic is looking for *is* a mental state).

## 2.1 Kripke's version of the paradox

Now, Kripke develops the paradox as follows. Suppose *S* is engaged in a computation he has never before carried out, using higher numbers than he has ever used, or anyone else for that matter. Let's suppose for simplicity's sake, following Kripke, that these numbers are 57 and 68. *S* performs the calculation and obtains the (correct) answer: 125. This answer is correct in two senses, both that $57 + 68 = 125$ is the correct arithmetical sum of these two numbers, but also in the 'metalinguistic' sense that *S* intended to use the symbol '+' in such a way that it referred to the addition function and no other function (Kripke 1982, p. 8).

*S* then encounters a 'bizarre sceptic' who challenges him to say what fact about him determines that he didn't use the symbol '+' in the past according to the following rule:

$$a \oplus b = \begin{cases} a + b & \text{if } a, b \leq 57 \\ 5 & \text{otherwise.} \end{cases}$$

This function is completely consistent with everything *S* has done up to now and

2. What I intend is something quite similar to what Martin Kusch has called *meaning determinism* (Kusch 2006). The reason I choose not to use that label is that I think it is a truism that meaning determines correctness conditions. The question is not: how does meaning determine correctness?, but rather: how can a *fact* determine that I meant anything in the first place? This is reflected in Wittgenstein's discussion, which is aimed at showing that a state of mind of a person is not necessary and sufficient for her to have meant something.

the sceptic's challenge is for us to find some fact in virtue of which *S* meant addition in the past and not this deviant function (which Kripke calls 'quaddition' or 'quus'), and by parity of reasoning, a fact that rules out any other function consistent with his practice up to now other than 'plus'. If there is no such fact, the sceptic claims, *S* cannot mean anything by his utterances since anything *S* might do falls under *some* concept, however deviant, and since there is nothing to decide between the various different concepts, *S* cannot mean anything by his words.

This problem is not an epistemological one. The question is *not* about how we can know or be justified epistemically that we did indeed mean *addition* and not *quaddition*, but rather concerning the constitution of the meaning of our utterances. The factualist about meaning tries to find some *fact* in virtue of which *S* meant one thing rather than another, while the sceptic argues that there is no such fact, and hence no meaning. In a later section, I will take a closer look at the notion of 'fact' involved, but for now, we just take it as given that we know what this means.

Kripke's argument is presented as a sceptical paradox, the conclusion of which is that there is no such thing as meaning—no such thing as *S* meaning *p* by his utterance 'p'. The character of 'the sceptic' demands a fact that determines what we mean by our utterances and claims that without such a fact, we cannot mean anything by our words. This of course threatens to make the sceptic's own words meaningless, and hence the development of the paradox itself incoherent: the sceptic is in danger of sawing off the branch on which he is sitting. For this reason, Kripke develops the paradox as a matter of finding a fact in the *past* that determines what *S* meant *then*. Kripke writes:

> The ground rules of our formulation of the problem should be made clear. For the sceptic to converse with me at all, we must have a common language. So I am supposing that the sceptic, provisionally, is not questioning my *present* use of the word 'plus'; he agrees that according to my *present* usage, '68 + 57' denotes 125. Not only does he agree with me on this, he conducts the entire debate with me in my language as I *presently* use it. (ibid., p. 11-12)

The sceptic, Kripke says, only questions whether my present usage conforms to my intentions in the past—if I meant *plus* in the past by my use of the word 'plus', then

I should say '125' *now*, and not '5'.

This is, however, a mere rhetorical device on Kripke's part. Later in the development of his dialectic, he makes the point that if there cannot be any fact in the past about what *S* meant by his utterances, then there cannot be any such fact in the present either (Kripke 1982, p. 21). There is therefore no restriction to past facts in Kripke's dialectic and Kripke's sceptic does in fact allow all facts to be considered—including any facts about *S*'s mental state, past *and* present.

We do not, however, need to be worried about incoherence. As Anandi Hattiangadi points out (Hattiangadi 2007, p. 21), Kripke's sceptic does not make any *a priori* argument for why his challenge cannot be met—rather, his argument proceeds by elimination.[3] Kripke considers and rules out a large variety of possible replies to the sceptic: that *S*'s mental states are what determines what he meant, that *S*'s dispositions are what determine meaning, that *S* follows a particular algorithm, that meaning is a primitive and cannot be given further analysis, platonism, and many others. Kripke's sceptic does nowhere argue that his challenge cannot be met at all, but rather argues that all the plausible options so far proposed have failed. His confidence in the paradox's conclusion is confidence that nothing better will be produced.

For that reason, it is not necessary to understand Kripke's argument as a sceptical argument at all (nor Wittgenstein's, for that matter). We can instead view it as as a challenge to give some account of the constitution of meaning on the one hand, and a powerful argumentative strategy to rule out many of those accounts that we might otherwise have found most intuitive and attractive, on the other. Viewed in this light, we do not need to worry about incoherence, since this question can be put to us without the conclusion that meaning is impossible, and thus the sceptical problem becomes, as Hattiangadi puts it,

> really just the hoary old problem of intentionality—of developing a theory that tells us what some linguistic or conceptual token is about, what it represents. (ibid., p. 211)

This is how I will understand Wittgenstein's rule-following considerations as well,

3. Although, it should be noted that Hattiangadi believes that such an argument can be constructed from the sceptic's arguments, by focusing on the normativity of meaning.

as an argument intended to clear away meaning mentalism, a view Wittgenstein otherwise finds attractive and important, so that a better view can replace it. That is to say, we should view the argument that leads to the paradox as a *reductio ad absurdum* of a certain intuitive picture of meaning, but not of the notion of meaning more generally. This is of course in line with the reading of *PI* §§186–242 outlined in the previous chapter, as well as Wittgenstein's emphasis on the connection between the concepts 'understanding', 'intention' and 'meaning'—both in *PI* and in the *Lectures*.

However, since Wittgenstein's target in the *Philosophical Investigations* is mostly meaning mentalism—the notion that meaning (and intention and understanding) is a mental state, it is useful to examine Kripke's arguments against a more general version of that view. In what follows, I will not focus on the question of how best to read Kripke nor the question of whether Kripke gets Wittgenstein right, but rather examine his arguments as the relate to Wittgenstein's version of the paradox, mostly to get clearer on what such a solution to the paradox requires.[4]

### 2.1.1 The argument from normativity

Most of Kripke's attention is devoted to answering various dispositionalist accounts of meaning whereby *S* means *addition* by '+' if and only if *S* is disposed to answer with the sum of two numbers, and not their quum. In a later section, I will focus on such accounts and give some of these arguments, but for now I want to discuss Kripke's most important argument against dispositionalist accounts, the argument from normativity: if one means something by a term, then one *should* or *ought* to use it in a specific way. The following two passages are instructive:

> A candidate for what constitutes the state of my meaning one function, rather than another, by a given function sign, ought to be such that, whatever in fact I (am disposed to) do, there is a unique thing I should do. Is not the dispositional view simply an equation of performance and correctness? (Kripke 1982, p. 24)

4. I do think, however, that Kripke's interpretation of Wittgenstein is better than he is given credit for—at least if one does not read Kripke as claiming that Wittgenstein is actually developing a sceptical argument.

> Suppose I do mean addition by '+'. What is the relation of this supposition to the question how I will respond to the problem '68+ 57'? The dispositionalist gives a descriptive account of this relation: if '+' meant addition, then I will answer '125'. But this is not the proper account of the relation, which is normative, not descriptive. The point is not that, if I meant addition by "+', I will answer '125', but that, if I intend to accord with my past meaning of'+', I should answer '125'. (Kripke 1982, p. 37)

Earlier, he had written that in the end, "almost all objections to the dispositional account boil down to this one" (ibid., p. 20).

The most uncontentious and paradigmatic examples of normative statements are those of ethics. For instance, the statements "Thou shalt not steal" or "Be kind to animals" tell us what we should or ought to do (or not do) without any appeal to our desires or objectives. They are *categorically* normative, or as we could also say, they contain 'genuine oughts'. If they are true, then if I fail to act in accordance with them, I will have done something wrong, and not merely failed in reaching my goal or objectice: they tell me what values or objectives I *should* have. The most obvious way to read Kripke's slogan that meaning is normative is to have read it in this way: If *S* means *F* by *w*, then *S should* or *ought* to apply *w* to all and only things that are *F*, where 'should' is read in the categorical or 'genuine' (albeit semantical, but not ethical) way (see e.g. Whiting 2007; 2016, for a defence of this view).

In the two passages above, however, Kripke rather seems to be making a connection between statements having correctness conditions—our principle (C)—on the one hand, and meaning being normative on the others. This connection comes out in an influential interpretation of Kripke, given by Boghossian:

> Suppose the expression "green" means *green*. It follows immediately that the expression "green" applies correctly only to *these* things (the green ones) and not to *those* (the non-greens). The fact that the expression means something implies, that is, a whole set of *normative* truths about my behaviour with that expression: namely that my use of it is correct in application to certain objects and not in application to others. […] The normativity of meaning turns out to be, in other words,

> simply a new name for the familiar fact that …meaningful expressions possess conditions of *correct use*. (Boghossian 1989, p. 513)

Crispin Wright gives a similar gloss on what the normativity of meaning amounts to:

> Meaning, it is again platitudinous to say, is normative: it is because statements have meaning that there is such a thing as correct, or incorrect, use of them. (Wright 2001a, p. 276)

Some philosophers have argued that this is not in fact a form of normativity at all. For instance, (Wikforss 2001) argues that the fact that the word 'horse' only applies to horses does not imply that one *ought* to apply the word only to horses—there might be other reasons not to do so, or even no reason one way or another. She furthermore points out that if I see a cow, and think that it is a horse, and then utter the statement 'This is a horse', I have done something incorrect but not violated a norm that is semantic in any way, since if I meant 'horse' by my utterance, I should say 'horse' (see ibid., p. 205–206). My mistake was not that I misspoke, I said what I intended. I merely misperceived. It is thus unclear how to get any kind of *ought* so directly out of the notion of meaning on the one hand and the idea of terms having correctness conditions on the other.

This corresponds to the distinction Anandi Hattiangadi has made between the *normative* proper and the *norm-relative* (Hattiangadi 2007). Meaning, on this view, is not 'normative' in the genuine or categorical sense at all, but 'governed by norms': meanings set a standard of correct and incorrect use, but have no prescriptive powers. This is also what Verheggen has called the 'trivial' sense in which meaning can be said to be normative: "According to the trivial sense, to say that meaning is normative is simply to say that linguistic expressions have conditions of correct application" (Verheggen 2011).[5]

But of course, nobody—not even Whiting—would argue that semantic normativity would be an overriding concern when evaluating what to do. If soldiers come knocking on the door in the middle of the night, looking for members of persecuted minority, one shouldn't say that they are hidden in the attic, because

5. See also Hattiangadi 2006; Glüer and Pagin 1999.

'they' refers to them and 'attic' means attic. Rather, Whiting says, we should think about them as *pro tanto* reasons—reasons telling us what to do if nothing else gives us an overriding reason (Whiting 2016).

In what follows, however, I will put this debate about normativity to one side and simply assume that what is at stake is this 'platitudinous' idea that our use of words has correctness conditions—that when we use words in any type of linguistic intercourse, some uses of that word will be right and some wrong. That is of course nothing else than our old principle (C). This way of understanding the normativity thesis is the weakest possible, so even if it turns out that meaning is normative in the more substantive, genuine sense, there would be no contradiction with this weak understanding, since both sides agree on at least this much.

One aspect of it has to do with the possibility of applying a word correctly or incorrectly to an object, giving rise to the following related principle (which avoids the use of the word 'ought'):

(C′) If *S* means *F* by his use of a symbol $\varphi$, then it is correct for *S* to apply $\varphi$ to an object *x* iff *x* is *F*.[6]

This formulation does not appeal to any kind of normativity *tout court* and while it is not in itself a transtemporal principle, it does put constraints on what *S should* do under the conditions specified, i.e. in order to have meant the same thing now as in the past. For example, the two-place addition function takes infinitely many values and if *S* meant to denote this function by use of the symbol '+' in his utterance of 23 + 57 = 80, then he is doing something incorrect, by his own lights, if he later says that 68 + 57 = 5. In other words, what *S* meant in the past, determines what he must say in a future case *in order to have meant the same thing as in the past and be correct about that future case*—when *S* meant addition by his use of the symbol '+' in the past and means *the same thing* now, in a future case, there is only one correct answer.

This merely follows from the fact that (at least some of) our use of words have correctness conditions which settle infinitely many cases: if *S* intended (or meant)

6. See Hattiangadi 2006; Glüer and Wikforss 2009; 2018, for a similar principle. This way of giving (C) does not fit nicely with Kripke's example, which concerns arithmetical concepts, not concepts that classify objects, but it generalises easily enough.

to refer to the addition function when he used the symbol '+' at $t_1$, the very fact that the addition function settles infinitely many sums entails that if *S* meant the same thing at $t_2$, namely to refer to the addition function by his use of the symbol '+', his meaning addition at $t_1$ had already established the correctness conditions for his use of '+' at $t_2$—that is what meaning the *same* thing at $t_1$ and $t_2$ *is*. The factualist about meaning maintains that there is some *fact* about *S* which makes it the case that he meant the same thing at $t_1$ and $t_2$, and in particular, the dispositionalist argues that this fact (or set of facts) is *S*'s dispositions regarding the use of '+'.

I will remain agnostic about whether or not this way of looking at things should properly be described as involving an appeal to the 'normativity of meaning', as it is not meant to imply anything about what *S* should do *simpliciter*, since *S* has no obligation to use '+' in one way or another, and the 'ought' involved is merely hypothetical (not a 'genuine' ought): it is only by having the intention of expressing the same thing as in a past case that *S*'s ought to use '+' in a particular way in a new case in the future, given the correctness conditions of '+' (see e.g. Hattiangadi 2006, on this point). It is also not committed to *S* having *pro tanto* reasons to say one thing rather than another.[7]

There are trivial ways to meet (C) and (C′), however, and we should also rule out accounts that judge *any* utterance *S* might ever make as being correct—namely, relative to *some* rule. In other words, an acceptable solution must not say that there are never any *incorrect* utterances, only different kinds of correct ones. This is, quite roughly, a problem of how to distinguish between *S* incorrectly giving the reply '5' to an addition problem or correctly giving that reply to a quaddition problem (and is sometimes called 'the problem of error'). A proposed solution must not say that whenever *S* makes—what we would call—a mistake in adding, e.g. saying that 68 + 57 = 5, he is actually carrying out a different calculation, namely quadding, where he does the right thing. Every purported error *S* would make on such an account would then simply indicate a difference in meaning, since that is what *S* did, and thus what he meant. In other words, *S*'s utterances must be able to be incorrect—*tout court*.[8]

7. This way of understanding the argument from normativity could be understood as expressing "not claims about what the subject ought to do but rather claims about what the subject is committed to doing" (Millar 2002).

8. See Wikforss 2001, for a similar discussion.

Similarly, we should also be able to explain mistakes that are not linguistic in nature—for there is a way of using a word correctly, while misapplying it. For instance, if *S* means *flu* by his utterance '*A* has the flu', then *S* has made a mistake if *A* does not have the flu, but this mistake is not necessarily a semantic mistake. If *S* believes that *A* has the flu or intended to mislead, then *S* has said what he intended, and so his use of words was correct after all. He was just mistaken about the facts or lied. To say that meaning is normative, then, is to say that there is some standard that *S* adheres to in his use of words—that there is a way of using words correctly or incorrectly.[9]

On the way I read the argument from normativity, then, the relation Kripke speaks of between meaning and intention, on the one hand, and future action on the other, should be understood in terms of meaning setting a standard of correct use. The skeptical demand is for some fact about *S* or his environment that can show that *S* is conforming to some such standard and how that standard is constituted—something that constitutes *S*'s meaning *x* rather than *y* and has the property that it distinguishes between correct and incorrect uses. In the example above, *S is* adhering to such a standard, despite his mistake.

**Justification and guidance** In giving his argument from normativity against dispositionalism, Kripke sometimes speaks of the meaning justifying *S*'s utterances. For him, the dispositional reply really "ought to appear to be misdirected, off target" (Kripke 1982, p. 23) because the sceptic has called into question *S*'s *justification* for saying '125' rather than '5' when asked about the sum of 57 and 68. If we accept the dispositionalist account, Kripke thinks, *S*'s response is no better than "a stab in the dark", as the dispositionalist reply only tells us what *S* is disposed to do now, and perhaps was disposed to do in the past. This is no good, Kripke thinks:

> Well and good, I know that '125' is the response I am disposed to give (I am actually giving it!), and maybe it is helpful to be told - as a matter of brute fact - that I would have given the same response in the past. How does any of this indicate that - now *or* in the past - '125' was an answer justified in terms of instructions I gave myself, rather

9. Thanks to Alan Millar for this example.

> than a mere jack-in-the-box unjustified and arbitrary response? (ibid., p. 23)

There are two aspects to this worth noting. First, Kripke seems to be saying that even if we had enough dispositions to cover all cases, that wouldn't be enough, since there wouldn't necessarily be anything that justified that response as the right response—what seems to be missing is that *S*'s utterances adhere to some standard of correct use in the right way (however that is correctly specified).

The second is that *S*'s meaning *addition* by '+' seems to be guiding *S*'s use of language. We feel that if we have grasped a rule like +2, then the rule and our grasp of it is guiding our application of it into new and new cases. This, we might say, following Kripke, marks the distinction between "someone who computes new values of a function and someone who calls out numbers at random" (ibid.). But how can the rule guide me if any fact that I might reasonably have access to—i.e. my mental state—underdetermines which rule I am in fact following? Kripke writes:

> Sometimes when I have contemplated the situation, I have had something of an eerie feeling. Even now as I write, I feel confident that there is something in my mind – the meaning I attach to the 'plus' sign – that *instructs* me what I ought to do in all future cases. I do not *predict* what I *will* do [...] but instruct myself what I ought to do to conform to the meaning. [... ] But when I concentrate on what is now in my mind, what instructions can be found there? (ibid., p. 22)

Famously, however, Wittgenstein argued against both notions. He constantly points out that justifications have to come to an end somewhere—if I'm asked why I wrote down 1004 after 1002, then I cannot appeal to my interpretation of the rule without leading to a regress of justifications: if I interpreted the rule in *this* way, then my interpretation again needs to be interpreted, and so on (*PI*, §217). He also famously argues that when we obey a rule, we do not choose, but do so *blindly* (*PI*, §219).

These claims are of course controversial and unclear, like so much else. Nevertheless, a solution to the paradox needs to be sensitive to these issues.

### 2.1.2 The 'sceptical' solution and non-factualism about meaning

Kripke distinguishes between two different ways of responding to the paradox: 'straight' solutions and 'sceptical' ones. The first kind of response involves denying that the sceptic's reasoning is cogent, perhaps by by denying one of his premises or producing a meaning constituting fact that we have hitherto overlooked. The dispositional account is a prime example of a straight solution, which I will discuss in the next section. A sceptical solution, on the other hand, is one where the sceptical reasoning is accepted and an argument is given to the effect that the area of discourse that the sceptical paradox threatens does not need the kind of justification that the sceptic had assumed—in our case, facts (Kripke 1982, pp. 66–67).[10]

Kripke argues that Wittgenstein's response to the paradox is a sceptical solution. For him, Wittgenstein does not give a new species of fact that the sceptic overlooked or deny any of his premises. At the same time, however, he does not wish to deny that people speak of others having meant such-and-such and that they are perfectly right in doing so (ibid., p. 69). Therefore, instead of specifying the truth conditions under which a statement of meaning is true, we should rather look at the circumstances under which such ascriptions of meaning can be legitimately asserted and what role such ascriptions have in our linguistic intercourse—we, as Kripke puts it, should

> widen our gaze from consideration of the rule follower alone and allow ourselves to consider him as interacting with a wider community. (ibid., p. 89)

If we do so, we will find that, instead of failing to find anything about *S* that constitutes his meaning, we will see various different circumstances under which meaning ascriptions are in fact asserted. Those, "assertability conditions", as Kripke calls them, are such that

> *Jones* is entitled, subject to correction by others, provisionally to say, "I mean addition by 'plus'," whenever he has the feeling of confidence – "now I can go on!" – that he can give 'correct' responses in new

10. Of course, if we agree that we shouldn't understand the paradox as a genuine paradox, but rather as a challenge to solve a philosophical problem, the distinction between a sceptical solution and a straight one becomes irrelevant. Here I merely follow Kripke.

> cases; and *he* is entitled, again provisionally and subject to correction by others, to judge a new response to be 'correct' simply because it is the response he is inclined to give. (ibid., p. 90)

Likewise, other members of the community will judge Jones by their own standards, and so on. In general, the sceptical solution holds that for *S* to mean *p* by his utterance of 'p' is for *S* to exhibit "sufficient conformity, under test circumstances, to the behavior of the community" (ibid., p. 96).

This solution is meant to replace our quest for a *fact* that constitutes *S*'s meaning. For Kripke, what makes us look for such a thing is a particular picture of meaning that Wittgenstein is trying to dislodge, that of truth-conditional semantics. Here's Kripke's version of it:

> [A] declarative sentence gets its meaning by virtue of its truth conditions, by virtue of its correspondance to facts that must obtain if it is true. For example, "the cat is on the mat" is understood by those speakers who realize that it is true if and only if a certain cat is on a certain mat; it is false otherwise. The presence of the cat on the mat is a fact or *condition-in-the-world* that would make the sentence true (express a truth) if it obtained. (p. 72. ibid., Emphasis mine.)

Kripke goes on to point out that if that is the general picture one has of the meaning of propositions generally, the idea that statements of meaning (i.e. the statement that *S* meant *p* by 'p') are likewise true in virtue of some fact (or as Kripke puts it, 'condition-in-the-world') that obtains becomes "not only natural but even tautological" (ibid., p. 73).

The idea, then, is to replace truth-conditional semantics by another picture of meaning provided by the sceptical solution. Kripke writes:

> If Wittgenstein is right, we cannot begin to solve [the sceptical paradox] if we remain in the grip of the natural presupposition that meaningful declarative sentences must purport to correspond to facts... The picture of correspondance-to-facts must be cleared away before we can begin with the sceptical problem. (ibid., pp. 78–79)

There is some controversy in the literature about how to understand this rejection of truth-conditional semantics and the notion of 'fact' involved, and hence what exactly Kripke intends the sceptical solution to do. Most commonly, it is seen a form of projectivism, whereby we

> we project an attitude or habit or other commitment which is not descriptive onto the world, when we speak and think as though there were a property of things which our sayings describe, which we can reason about, know about, be wrong about, and so on. (Blackburn 1984a, p. 170–71, as cited by Kusch 2006)

On this construal, the non-factualist about meaning holds that attributions of meaning are not, as Kusch puts it, "fact-apt" (ibid., p. 148). The non-factualist about meaning therefore denies that we can infer from a statement such as '*S* meant *p*' the corresponding statement "it is a fact that *S* meant *p*"—just as we wouldn't make the transition from "boo murder!" to "it is a fact that boo murder!".

Others, for instance, Scott Soames (Soames 1998) see attributions of meaning as performative speech acts, so that an utterance of "S means *p*" should be seen as on a par with utterances that do not express propositions at all, for instance the utterance "I hereby pronounce you husband and wife" as uttered by the right person in the right circumstances. Utterances of this kind should rather be seen as taking "Jones into one's linguistic community, to certify him as a competent user of '+' and to license him to use '+' to do what we call 'adding' " (ibid., p. 322). In this case, it is not so clear that we couldn't make an inference from, e.g. "I hereby pronounce you husband and wife" to "it is a fact that I hereby pronounce you husband and wife".

Several authors, most prominently Crispin Wright (Wright 2001c) and Paul Boghossian (Boghossian 1989), have argued that non-factualism about meaning is after all incoherent. Wright, for instance, has argued that non-factualism about meaning would be unstable and lead to global non-factualism, while Boghossian claims that local projectivism is necessarily committed to two incompatible views of truth, and hence, in his view, a contradiction. Since then, however, some authors have argued that Kripke's Wittgenstein is actually not a non-factualist about meaning at all, but is rather taking aim at a particular conception of truth-

conditional semantics, one where 'fact' is understood in a robust or concrete way (Byrne 1996; Soames 1998; Wilson 1998; Kusch 2006).

On this view, there *is* a fact of the matter whether or not *S* meant *p* by his utterance of 'p', not in the sense of there being "a condition in the world" which is such that *S* meant *p* if and only if that condition obtains, but rather in some minimal way—perhaps by stipulating, on the model of minimalism about truth, that fact ascriptions such as "it is a fact that *S* meant *p*" are true if and only if *S* meant *p*.

I will not take a stand on this debate here, as a reading of Kripke, but rather argue that this reading fits quite well with the reading given of Wittgenstein's remarks in the last chapter. George Wilson, one of those commentators, has given the following reconstruction of Kripke's Wittgenstein's argument (Wilson 1998):

(1) The meaning of a statement *p* is given by its truth-conditions.

(2) Truth-conditions are facts or conditions-in-the-world.

(3) In particular, the meaning of a statement of the form 'S meant *p*' is given by its truth-conditions—some fact or condition-in-the-world obtaining.

(4) There are no such facts about *S*.

(5) Therefore, statements of the form 'S meant *p*' are meaningless.

For Wilson, we should see Kripke's Wittgenstein as offering a *reductio ad absurdum* of (1): Statements such as '*S* meant *p*' are not meaningless, and hence (1) is in fact false. Wilson then argues that if we reject (1), there is no reason for us to go in for meaning factualism.

The trouble with Wilson's interpretation of Kripke along these lines is that is the sceptical solution simply ceases to be one: the difference between a sceptical solution and a straight one is meant to be that in the former case the reasoning of the sceptic is accepted as being unanswerable, but in the latter it is rejected as being faulty. On Wilson's reading, Kripke's Wittgenstein rejects a fundamental premise of the sceptic, namely (1) that meaning is given by truth conditions, and subsequently rejects the sceptic's conclusion that there is no such thing as meaning (on this point, see Miller 2002). True, the sceptical solution is meant to save

the appearances of our meaning stating discourse, but not by rejecting any of the sceptic's premises.

I will not be particularly bothered about this here. It seems to me that even if Wilson is wrong about Kripke, it might be useful to examine Wittgenstein's arguments against meaning mentalism rehearsed in the last chapter in this light (even if Wittgenstein nowhere makes this argument explicitly).[11] On the view criticised there, *S* meant *p* by his utterance of the sentence 'p' if and only if *S* was in some appropriate mental state that characterised his meaning. Such a mental state, it would be easy to argue, is a fact or a condition-in-the-world and so, the proposition that *S* meant *p* by his utterance of 'p' gets its meaning in virtue of its truth-conditions, namely *S*'s mental state. The argument would then run in the same way, except with the added assumption that the fact about *S* that gives the meaning of '*S* meant *p*' is a mental state.

On this reading, Wittgentein's denial that there is some fact, e.g. a mental state, such that *S* meant *p* if and only if that fact obtains, does not imply that there is no fact of the matter whether or not *S* did in fact mean *p*, simply because it doesn't follow that because truth-conditions are facts, then facts must be understood robustly or as 'conditions-in-the-world' (or however we want to cash that out). In other words, 'it is a fact that *p*' does not imply that *p* is a 'condition-in-the-world' and the statement 'it is a fact that *S* meant *p*' does not imply that there is some fact about *S* which makes this the case.

How might we try to motivate this claim better? Wilson states the non-factualist thesis in the following way:

(NF) There are no facts about a speaker in virtue of which ascriptions of meaning—even among those that are fully warranted by all our usual criteria—are correct.

In this formulation, the clause "those that are fully warranted by all our usual criteria" is doing most of the work, as it is intended to make room for fact-stating discourse that is not simply about 'conditions-in-the-world'. If we make such room, Wilson argues, Wittgenstein does not have to accept (NF) and moreover,

11. But see Chapter 3 for more discussion.

this is not in any contradiction with his rejection of (1)—that only follows if "having classical truth conditions and purporting to describe facts are taken to be one and the same" (Wilson 1998, p. 114). Wilson concludes that "there is no reason for Kripke's Wittgenstein to do so" and that presumably,

> using the resources of the 'sceptical solution', Kripke's Wittgenstein will want to offer his own account of what uses of language can count as fact describing. And, I can not see that anything in the 'sceptical solution' forecloses that option for him. (ibid., p. 114)

Wilson offers the following parable in order to draw out the differences between (NF) and (1) (ibid., p. 117): Suppose two philosophical opponents, *A* and *B* agree in the following proposition:

(i) The numerals are names of the natural numbers.

They disagree, however, in the content of the proposition. *A* thinks that (ii) the numerals are proper names of objects or entities, the natural numbers, while *B* thinks that this is misguided. For *B*, (i) is used to

> register certain fundamental facts about the surface grammar of the numerals, and, especially, facts about their distinctive, non-proper-naming use in language, i.e. their use in counting, in arithmetic calculations, and in sentences that record the results of these countings and calculations. (ibid., p. 118)

Now suppose that *B* comes up with a Benaceraff-style argument, aiming to show that (iii) there can be no facts that establish that a particular entity is named by a numeral 'n' (see Benacerraf 1973). Given *A*'s reading of (i) as (ii), *A* will now reformulate this sceptical conclusion as (iv) there are no facts that establish what the numeral 'n' names, and conclude that while the numerals are proper names, they do not name anything at all (Wilson 1998, p. 118).

Wilson, however, argues that since *B* does not accept *A*'s characterisation of (i) as proper names of entities, she likewise does not have to accept the transition from (iii) to (iv)—and in fact, she will judge that (iii) is true and (iv) false. For instance, *B* might hold that there are certain facts about our mathematical practice

that make it the case that our numeral '1' names the number 1 and we can, Wilson says, grasp truths of this kind when we've understood the concept of 'naming a number'. This allows her to further block the transition from (iii) to (iv′):

(iv′) There are no facts in virtue of which statements of the form 'Numeral 'n' names the number N', even when warranted by all our usual criteria, are correct

which is the analog of (NF) in this story.

It is of course no coincidence that one of the characters in Wilson's parable sounds a lot like Wittgenstein (and the other like a mathematical platonist) and while Wilson does not tell us much about *how* we should understand these facts about our practice or what indeed it is to understand a concept such as 'naming a number', this story does show that the move from the rejection of truth-conditional semantics to non-factualism about meaning is not as automatic as it would first seem. Wilson's point is that there is no reason to reject the possibility of giving a different account of meaning, one that does not require truth-conditional semantics, but at the same time does not commit itself to non-factualism about meaning—even if Kripke's sceptical solution fails.

## 2.2 The inadequacy of dispositionalist accounts

A large part of Kripke's discussion is intended to deflect dispositional accounts of rule-following and meaning. The following is Kripke's first pass at a definition of what such accounts amount to:

> To mean addition by '+' is to be disposed, when asked for any sum '$x$ + $y$' to give the sum of $x$ and $y$ as the answer... (Kripke 1982, p. 27)

Apart from the argument from normativity, canvassed above, Kripke has two main arguments against dispositionalism which will be relevant for us. The first argument is simply that it is false that we are disposed to give the sum of any two numbers when queried: *S*'s doesn't actually have the disposition to give a reply in infinitely many cases and to mean addition by the symbol '+', *S* would have to have the disposition to give a reply in every case. Kripke writes:

> …some pairs of numbers are simply too large for my mind—or my brain—to grasp. When given such sums, I may shrug my shoulders for lack of comprehension; I may even if the numbers are large enough, die of old age before the questioner completes his question.(ibid., p. 27)

That is to say, it's quite questionable to suppose that *S*, given his finite human nature, could ever have all the dispositions necessary to have meant an infinite function by a finite symbol. There would be some correct sums that *S* has no dispositions for, if we assume this model of meaning.

The second argument is aimed at a possible repair of this naïve version of the dispositional account. The idea is to complicate the notion of disposition such that *S* meant addition by '+' if *S*, *ceteris paribus*, would have responded with the sum of two numbers $n + k$ when queried in hypothetical situations where *S* has the necessary abilities and time to give a reply. Kripke's reply is that such an fix would only work if we would already assume that *S* would reply with the sum of $n + k$ and not their quum, an assumption we cannot make without begging the question. The trouble is, how can we spell out this *ceteris paribus*-clause without simply assuming *S* is adding and not quadding? If *S* is an adder, it is true, he will, *ceteris paribus*, add the two numbers, but if he is a quadder, he will then quadd them, and nothing about *S* so far makes this distinction.

A more promising iteration of the naïve dispositionalist account just sketched is the algorithmic account—whereby our dispositions are best seen as 'internalised instructions' on how to carry out an algorithm that corresponds to addition. On this picture, *S* means 'addition' by his corresponding utterances if *S* is disposed to carry out this algorithm. Kripke gives one example of such an algorithm for adding two numbers, $x$ and $y$:

> Take a huge bunch of marbles. First count out $x$ marbles in one heap. Then count out $y$ in another. Put the two heaps together and count out the number of marbles in the union thus formed. The result is $x + y$ (ibid., p. 15).

To mean addition, then, is to be disposed to carry out this algorithm when using the symbol '+'.

Kripke's argument against this reply are in effect that since we have only applied the algorithm in finitely many cases in the past, the sceptic can reinterpret *S*'s use of *count* in such a way that it fits with his use in the past but deviates from counting proper. This deviant interpretation would contaminate upwards, and make our use of addition deviant as well. Likewise, if we were to find some algorithmic explication of *count*, we could do the same thing again, all the way to the most basic concepts, whichever they are, and deviantly reinterpret them.

In Chapter 7, I discuss the algorithm reply further.

### 2.2.1 Warren's 'canonical' dispositional account

A recent dispositionalist account that takes the algorithmic reply as its starting point is that of Jared Warren (Warren 2018). It is instructive to see why Warren's account does not work, because his approach essentially relies on complicating the notion of disposition in such a way that it can handle infinitely many cases and have the structure required. It cannot, however, I will argue, answer the crucial Wittgensteinian question of what constitutes the meaning of words—i.e. it cannot account for the correctness conditions of meaning.

In his paper, Warren gives a good dramatisation of what were are looking for to solve the problem extensionally by supposing that *S* has a *disposition table* from which we can read *S*'s dispositions (ibid., p. 3). Such a table represents, for each ordered pair $\langle n, k \rangle \in \mathbb{N} \times \mathbb{N}$, the numeral '*m*' which *S* is disposed to give in response to queries about simple sums of the form "what is $n + k$?". If some such table *T* contains a numeral '*m*' for every pair of natural numbers (i.e. *T* is defined by a function from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{N}$), we say that it is a *full table*. Likewise, if *T* is such that whenever a pair $\langle n, k \rangle$ in *T* is associated with some numeral '*m*', there is another table *U* such that $\langle n, k \rangle$ in *U* is also associated with '*m*', we say that *T* is *compatible* with *U*. Simply put, *U* might either be a subtable of *T* or $T = U$, if *U* is a full table. If *S* means addition by '+', then *S*'s disposition table is compatible only with the addition table and no other full table.

Warren agrees that the algorithmic reply cannot work as Kripke specifies it, but thinks that we can give a similar dispositional reply which does not require "any appeal to internal instructions" (ibid., p. 6). This would naturally solve the problem,

he claims, since there wouldn't be anything for the sceptic to deviantly reinterpret. Warren's own account is an elaboration of attempted solutions by Simon Blackburn and Tomoji Shogenji (Blackburn 1984b; Shogenji 1993). The 'key idea' Warren borrows from Blackburn and Shogenji is making a distinction between simple and complex dispositions as follows (Warren 2018, pp. 6–7):

(Simple) *S* has a simple disposition to $\varphi$ in situation *C* iff *S* $\varphi$s in *C* directly, not by way of performing any intermediary actions or activities.

(Complex) *S* has a complex disposition to $\varphi$ in situation *C* iff S $\varphi$s in *C* as a result of performing some intermediary actions or activities that *S* is disposed to undertake in *C*.

This distinction is supposed to be made at the 'level of conscious actions', according to Warren. A simple disposition is brute and not the outcome of any operation or multi-step process of which *S* is conscious (even if such processes might take place subconsciously, e.g. in the brain). These simple dispositions are the result of training and repetition, and "encoded in instinct and unconscious habit" (ibid., p. 7). In contrast, complex dispositions are the result of some conscious process or operation which *S* performs in the world—like Kripke's counting algorithm. When asked about some sum for which *S* has no simple disposition, e.g. $n + k$, he might have the complex disposition to add single digit numbers, perform the carry operation if the sum is larger than 9 and so on. *S* can thus have a complex disposition to reply with the sum of $n$ and $k$, even if he has no such simple disposition. This distinction, however, means that there is a kind of hierarchy of dispositions where the simple dispositions are on bottom, forming the more complex dispositions on top.

Warren points out that this view does not depend on any internalised instructions (or at least not obviously) but merely dispositions to act in certain ways, and thus there is nothing for Kripke's sceptic to reinterpret:

> In any case, sceptical interpretations of any explicit instructions Ludwig [i.e. my *S*] happens to give himself miss the point of this response: having a complex disposition to reply with the sum, by way of execution of the addition algorithm, is not "just more theory" for the sceptic to interpret. (ibid., p. 7)

Warren is correct that by dropping the appeal to internalised instructions, dispositionalist accounts are not as " obviously vulnerable"[12] to Kripke's arguments, but as we will see later in the chapter, that is not enough, as the issue is not that there is always more theory to refer to, but how the extension of our most basic concepts and rules gets fixed in the first place—what constitutes meaning.

For now, however, there is a more immediate problem for Warren. In order to deal with truly astronomical numbers on this account, we still need some kind of *ceteris paribus* clause—which, as we know from Kripke, spell trouble for dispositional accounts. This is because *S* might not have any complex dispositions for such numbers, since he might either die or give up before ever executing the algorithm. If there is a problem about *S* being simply disposed to reply with the sum of astronomical numbers, then surely there is also a problem of *S* having a complex disposition to go through the process of calculating an even larger number: nobody is disposed to go through a calculation which takes longer than their lifespan, for example.

Warren seeks to solve this problem by making a further distinction between dispositions which is intended to capture the way in which the "unbounded execution of the process is already encoded in Ludwig's starting dispositions" (Warren 2018, p. 8).

The idea is to appeal to the fact that *S*'s dispositions are *linked*. When *S* has completed a particular step of the algorithm, *S* is disposed to execute the next step if and only if the previous step has been completed. Warren writes:

> Ludwig might enjoy putting together bookshelves, following an implicit algorithm for doing so that involves working from the bottom to the top to install the shelves. He might then be disposed to first install the bottom shelf, and then, with the bottom shelf installed, to install the next bottommost shelf, and so and so forth. These dispositions of his are linked. (ibid., p. 8)

Based on this idea, Warren makes a further distinction between dispositions is as follows:

12. Warren cites the authority of Martin Kusch, a stalwart defender of Kripke's paradox, on this point. However, Kusch's emphasis is clearly on the obviousness of the vulnerability, not the vulnerability itself. (Kusch 2006, p. 112).

(Singular) *S* has a singular disposition to $\varphi$ in situation *C* iff *S* has a simple (or complex) disposition to $\varphi$ in *C*.

(Composite) *S* has a composite disposition to $\varphi$ in situation *C* iff $\varphi$ing is (or would be) the output of the iterated application of S's linked simple (or complex) dispositions, in *C*.

The point is, Warren says, is that even if Ludwig lacks the disposition to reply with the sum of two astronomical numbers when asked, he would still be disposed to execute the first step of the algorithm for addition, and then for each *particular* step *n* of the algorithm to execute that step and move on to step $n + 1$. This, he adds, does not mean that he has the complex disposition to execute all the steps of the algorithm, since particular complex dispositions for astronomical numbers might not be present.[13]

In this sense, Warren's solution is a kind of induction (or recursion) of dispositions: Ludwig has the simple (or singular) disposition to carry out step 1 of the algorithm and the simple disposition to carry out step *n* and move to step $n+1$, for any *n*. It therefore follows that he has the composite disposition to give the sum of any two numbers, or so Warren claims. This way, we can fill out *S*'s full disposition table for '+' as follows: the entry in the table given by a row *n* and column *k* is the numeral '*m*' such that '*m*' is generated systematically from the dispositions *S* has to "sum single digit numbers, carry, and move on to the next step in the process *for each step*" (ibid., p. 9). Warren writes:

> This is a matter of him, for each step in the process, being disposed to perform step *k* [of the addition algorithm] and then move to and perform step *k+1*, not a matter of him having the conjoined disposition to perform step *k* in the relevant situation and also having the disposition to perform step $k + 1$. (ibid., p. 9)

The latter, Warren says, would suggest that there was no structure to *S*'s disposition, but there is: *S* is disposed to start with the first step of the algorithm, and then, when the first step is completed, he is disposed to move on and execute the next step.

13. Given Warren's example, we might therefore call this the IKEA-model of dispositions.

The idea, it seems to be, is that *S*'s dispositions are structured in the right way so that *S* can always generate more and more outputs *of the actual addition function* without having actually completed the previous steps—the problem that felled the previous iteration of this account. This idea, however, clever as it is, cannot work. To see why, notice that Warren's account implicitly relies on *S* following or being disposed to follow a general rule of how to move on to the next step, roughly of the form

(R) For all $k$, perform step $k$, then move to step $k + 1$.

This is of course not the only possible form of such a recursive rule, and while Warren describes his account without mentioning a rule, he is relying on something like (R) when he describes how *S*'s dispositions are linked: namely that if *S* is disposed to execute step $k$ of an algorithm and then to move on to step $k + 1$ of that algorithm, *S* is disposed to follow some rule like (R). This should of course not be surprising, since if *S*'s singular dispositions are linked to form composite dispositions, they must be linked in *some* way, and if his dispositions have a structure, they must have *some* structure. This structure is an infinite one, generated from *S*'s simpler, finite dispositions, and since it is infinite, it cannot be described without some generality or variables bound by quantifiers entering the picture somewhere.

As a consequence, Kripke's sceptic can always ask what fact about *S* makes it the case that the recursion rule *S* uses is indeed the recursion rule for addition, rather than quaddition? There is nothing about (R) which marks it out as the rule for addition, rather than the other infinitely many possible disposition tables. The rule is only the recursion rule for addition *if* we assume that the variable ranges over the correct disposition table, a question begging assumption, similar to the assumption other *ceteris paribus*-clause theorists have had. Warren simply assumes that the steps *S* is disposed to follow are the steps of an algorithm that generates the disposition table for addition, and this we cannot simply assume without begging the question.[14]

14. We could give other recursive rules for *S*. For instance, what fact about *S* makes it the case that he's not following the rule (R′): 'For all $k$, if $k < n$ perform step $k$, then move to step $k + 1$, otherwise give the reply 5'? However, we could always change (R) into (R′) by specifying a different domain of $k$'s.

The dispositionalist might give the following counter-argument to this objection: It is (trivially) true that *if S* were to be following the general rule resulting in the dispositional table for addition, then *S* would have the disposition to mean addition by his utterances. It then follows that if *S* was disposed to follow some rule like (R) which does not quantify over the steps of the addition algorithm, *S* simply didn't mean addition by his utterances. My objection, they might say, is merely tantamount to me saying that *S* couldn't have the right dispositions to add or that the mere possibility of a different set of dispositions shows that something is wrong. And there is just no reason to suppose that this is the case. Here's Warren:

> At this point, defenders of Kripkenstein may, in desperation, resort to expressing doubts that Ludwig really has the disposition to apply the addition algorithm for any pair of numbers. Perhaps, they may suggest, he is actually disposed to stop applying the procedure after the trillionth digit... [...] I admit that if Ludwig lacks these dispositions, he doesn't mean addition by "+", but why should we think that he lacks them? Kripke's worry that we lacked dispositions to reply with the sum for astronomical numbers was rooted in principled arguments showing that beings like us *couldn't* have simple dispositions for such cases, but here is nothing analogous here. This response is mere unprincipled gainsaying. (Warren 2018, p. 12)

This, however, is to miss the point. It is not merely that *S couldn't* have the right dispositions or that the mere possibility that he could have other dispositions shows that something is wrong. That was never Kripke's point to begin with, and his arguments about the difficulty of getting dispositions to have the right shape are really just a distraction from the real argument, that from normativity, That argument, as we saw above, really is about *S* adhering to some standard of correctness in is utterances—the very constitution of correctness conditions. Warren has not even begun to touch this problem.

That he does not can be seen immediately from how Warren sets the problem up for himself. Kripke's objection to the algorithm reply assumed that *S*'s had given himself some internal instructions that he himself needs to interpret in order to carry them out, leading the sceptic to strike elsewhere. Warren wants to find a

solution that does not need this assumption, and does so by casting himself as a theorist that describes *S*'s dispositions from without—in a metalanguage that takes all the relevant concepts for granted. The problem, however, was to account for how correctness conditions are constituted in the first place, and in explaining that, it is irrelevant to explain how *S*'s dispositions track those concepts.

We can perhaps best see why by asking how Warren himself might acquire a new concept—or indeed any speaker of Warren's metalanguage, namely us, the philosophers reasoning about *S*'s dispositions. Suppose Warren has seen a few examples of a colour he's never seen before, and says to himself: I shall call this colour 'bleikr'. On Warren's account, we should say that Warren has acquired the concept *bleikr* if and only if he has acquired the composite disposition to call bleikr things by the word 'bleikr'. But which things are that? Warren hasn't told us any story about how the correctness conditions of the term might be set up for himself in this case—and that strikes at the core of the problem: what picks out the correct sameness relation from the past to the future as correct? In this hypothetical example, Warren has seen a few examples and formed some dispositions, but correctness doesn't seem to have entered the picture at all.

If we suppose that our dispositions (i.e. those we have as speakers of the metalanguage) are a way out of this problem, we cannot do so without begging the question, since what we mean by a term such as 'bleikr' is precisely what determines what counts as bleikr (the correctness conditions of the concept *bleikr*) and so if our dispositions to use the term 'bleikr' determine what we mean by 'bleikr', anything we could possibly be disposed to do will be correct: our dispositions will be what we mean and what we mean determines what is correct. The dispositionalist account, as Kripke says, therefore confuses performance with correctness.

The fact that we do not know what 'bleikr' picks out on either side of the conditional draws this problem out—the relation between the word 'red' and the concept *red* is contingent and must have been set up somehow, from finite examples to indefinite correct uses. That is (one aspect of) Kripke's problem, and Warren offers no answer—and it is unclear how he ever could, given the role dispositions play in his account. For Warren, *S* means 'bleikr' by *bleikr* if and only if *S* has the full disposition table of the concept *bleikr*—this example shows that dispositionalist accounts cannot account for how a table gets picked out as being the table for *bleikr*

in the first place. It could be anything!

It might be objected here that I'm making things too hard for the dispositionalist, that I'm forbidding them the use of *any* concepts in the metalanguage—in which case we could not even begin to either express the paradox nor a solution to it. That is not my point, however—as the example about 'bleikr' shows. The point is rather that the real problem isn't about how *S* might mean something by a term, but how a standard of correctness is set up for our use of terms in the first place. It doesn't matter how complex our hierarchy of dispositions is, and saying that the recursive rule *R* combines our simple dispositions into the full dispositional table for addition only if the full dispositional table it generates is that for addition is trivial, and since we've only used that rule finitely many times, there can be no fact of the matter whether this is so.

Hence, there would be no use for Warren to dig in his heels and insist that *he* only means addition if and only if *he* is disposed to add, since he still hasn't told us anything about how the relation between the term '+' and *addition* was set up in the first place–i.e. which sameness relation from Warren's exemplars of addition his use of the word 'addition' picks out. It would be as if Warren were to insist that he means *bleikr* by 'bleikr', and so we should be able to tell what it means that *S* means *bleikr* by having dispositions to call bleikr things 'bleikr'. Thus, the dispositionalist account doesn't have any story to tell about how a finite set of exemplars could ever determine the correctness conditions for infinitely many occasions of use in the future—that is, pick out one sameness relation from the set of exemplars to novel cases.

If we could simply *stipulate* that our composite dispositions track the rule which is operative in our practice, namely *addition*, there would be no problem. The point is therefore neither desperate nor mere unprincipled gainsaying, but has to do with the very essence of the rule-following paradox: finding a fact about us that can give us correctness conditions for an indefinite range of uses of our concepts. Hence, even a dispositionalist account that manages to avoid Kripke's 'poverty of dispositions' arguments does not solve the problem, and that seems to be the best we can get out of such an account.

## 2.3 Problems with community solutions

Kripke's sceptical solution is a form of *community solution*—a solution that makes an essential appeal to a community of language users. The rough idea, as we saw from the sceptical solution, is that while an individual cannot follow a rule or make a meaningful utterance, we can evaluate their actions or utterances against the background of community practice and therefore get the correctness conditions that we want—or, perhaps in the case of the sceptical solution, an *ersatz* version which does the job well enough.

Most simply, the correct action in a given case on a community solution to the paradox would then be that which agrees with the community, and an incorrect one is one that is out of step with the community. David Bloor's expression of the idea is especially clear:

> Making a step in following a rule counts as a 'right' step, i.e. a genuine and successful piece of rule-following, if it is aligned with the steps everyone else, or nearly everyone else, takes.
>
> To make a 'wrong' move is ultimately to make a move that leads the individual along a divergent path. To be wrong is to be deviant, thought that isn't to say that 'wrong' *means* 'deviant'. (Bloor 1997, p. 16)

Such solutions can take both sceptical and straight forms. Perhaps the most prominent form of a straight solution is a dispositional community solution. whereby the dispositions of the community are taken to constitute the correctness conditions of a term: if *S*'s utterance agrees with the dispositions of the community, *S* was correct, and incorrect otherwise.

The following is Anandi Hattiangadi's criticism of the sceptical solution, which is fairly typical in this regard and generalises to most communitarian solutions, including dispositional community solutions:

> Any given individual's use of an expression is correct only if it is acceptable to the rest of the community. If the individual's use is unacceptable to the rest of the community, that use is incorrect. But the dispositions of the community taken together do not track an investigation-

> independent property either. Therefore, there is no possibility of mistake for the community as a whole. We may all be disposed to call some non-square things 'square'. (Hattiangadi 2007, p. 93)

The idea behind this criticism, it seems to me, is that just like the individual, the community taken as a whole has only calculated finitely many sums, and so there is nothing about the community's dispositions that determines whether the next calculation is correct or not, or perhaps rather nothing that determines whether the next step was the same action as the previous ones.

Crispin Wright has put the matter thus:

> The difficulty is to stabilise the emphasis on basic propensities of judgement against a drift to a fatal simplification: the idea that the requirements of a rule, in any particular case, are simply whatever we take them to be. (Wright 2007, p. 487)

By appealing to the dispositions of the community, either in a straight solution or a sceptical one, we've therefore simply moved the problem up a level: there are still going to be different sameness relations from the past to the future for the community as a whole, but since anything the community does is defined as correct, it is completely unconstrained in its practice.

In effect, this is just the old problem of error, moved up to the level of the community. Consider Boghossian's 'horsey-cow' case (Boghossian 1989, p. 173): Imagine that I stand in a field on a dark night and come across a cow which—unusually—looks quite like a a horse. While I typically do have a disposition to apply the term 'horse' to horses and the term 'cow' to cows, I'm also disposed to apply the term 'horse' to horsey-looking cows on dark nights. Furthermore, we can suppose that everyone in my community is also so disposed. And if so, then we cannot make any distinction between horses and cows on dark nights. Boghossian writes:

> The point is that many of the mistakes we make are *systematic*: they arise because of the presence of of features – bad lighting, effective disguises, and so forth – that have a generalizable and predictable effect on creatures with similar cognitive endowments. (ibid., p. 173)

And if that is the case, the communitarian cannot really say that I have made a mistake when I called the horsey-looking cow a horse, and must insist that 'horse' doesn't in fact mean *horse*, but rather 'horse or horsey-looking-cow-on-a-dark-night'—an intolerable conclusion.

The same would be the case for mathematical concepts. If the community solution to the paradox is right, then there could not be such a thing as a mistaken proof accepted by everyone, since by definition, if everyone accepted it, it would be correct. Nevertheless, our practice does make use of this distinction and it seems like a very repugnant conclusion to erase it. Ideally, what we'd want from a solution to the paradox, whatever it is, is that it is able to support a tripartite distinction between what *S* judges to be the case, what *S*'s community judges to be the case and what really is the case. Community solutions in general struggle with this problem.

There is at least one further problem, however. How does what the community does establish correctness at all? It cannot be that what is correct is so because *everyone* in the community does it that way, since that would mean that a single defection leads to there not being any such thing as correctness, as Bloor's exposition of the community view correctly rules out. But how does Bloor's caveat, that correctness is being in step with what "nearly everyone" does help? Is it just a sheer percentage that determines correctness, and if so, what is that percentage? Any such point would seem completely arbitrary and proponents of community solutions have not given any explanation of *how* the mere fact that more people do it one way rather than another make the former way correct—and just stating that there are in fact *much* more people in the first group doesn't seem to do the trick. It is unclear how the community solution actually achieves what it is purported to do.

In Chapter 4, I argue that what is missing is *structure* in our practice, and argue that we can advance a community solution that avoids the problems outlined here by imposing a structure on our practice.

# Chapter 3

# The relationship between meaning and rules

A widespread, almost orthodox, conception of language is that it is an activity *constituted* and *guided* by rules. By this, it is meant that the meaning of words are determined by the rules that apply to them and that the rule instructs speakers in how to use words. It is then supposed to follow that learning a language is a matter of *grasping* the rules governing the use of words and subsequently follow them when performing speech acts. This picture is meant to explain why some applications of words, i.e. when predicating something of an object, can be correct or incorrect: when the rule was followed correctly, the word was used correctly, and if the rule was broken, the word was used incorrectly, and it explains how we are able to speak a language at all: it is by grasping the content of the rule that we can apply our concepts in novel cases, as well as point to in justifying our performance and to criticise the performance of others.

So far, we've taken a look at a few different accounts of rule-following. In the background, I have not assumed a stronger principle than (C)—that there are correctness criteria for our use of concepts and the terms that refer to them. This principle is uncontroversial and I'm not aware of any philosophers that deny it. Often, however, philosophers conflate—rightly or wrongly—the presence of correctness conditions of a term with the idea that there is a *rule* for the use of the term. This gives rise to the following principle, equivalent to (C), if that assumption is

made:

(R) A term has meaning if and only if there is a rule for its use.

There are good reasons to accept this equivalency, above and beyond the picture just sketched—namely that rule-following and meaning share certain fundamental features that seem to be intertwined. For one, both rules and meaning have correctness conditions, i.e. it is both the case that when we mean something and when we follow a rule there is something which is correct and something which is incorrect. Secondly, both a rule and a concept have an indefinite (or even infinite) range of application and yet can only be learned or grasped by seeing a finite set of examples or receiving a finite amount of training. It would then seem that for any concept we can think of, there is a corresponding rule, namely the one that has the same correctness conditions and taught using the same examples and training.

Take for instance the concept *red*. It has some correctness conditions, namely that it is only correct to apply the term 'red' to red things. There is also a corresponding rule, albeit a trivial one: "Only apply the term 'red' to red things!" and if we have grasped this rule, whatever that means, it seems right to say that we have also grasped the corresponding concept. It is then not implausible to think that the correctness conditions of the word that refers to this concept, i.e. 'red', is determined by the rule, and indeed, many philosophers seem to think that having correctness conditions presupposes a rule: whenever there are correctness conditions, there is a corresponding rule which constitutes the correctness conditions.[1]

Notice however, that (R) does not itself imply that the correctness conditions for the use of the word 'red' are due to or constituted by the rule that links together the word and the concept, only that for every meaningful term there *is* a rule for its use. It might be that there is a third, and different, source for the correctness con-

1. See e.g. Wright 2007: "It is natural to think that in any area of human activity where there is a difference between correct and incorrect practice, which we achieve is (partly) determined by rules which fix what correct practice consists in, and which in some manner guide our aim." Wikforss describes the view thus:

> The idea is that in using my words I must be guided by a general rule, an 'inner instruction', telling me how to apply the word in the particular case. (Wikforss 2001, pp. 216–217)

ditions for the two—the meaning and the rule—and they they merely correspond in a trivial way (and indeed I will argue that this is the case in Chapter 4).

There is a different way, however, in which this principle does not seem to be strong enough. It is clear that merely acting in accordance with a rule is not enough to be actually following it. For instance, a dog might walk on the right side of a path, following its owner on a walk, but we wouldn't say that the dog is following the rule 'walk on the right side of the path!', even if that rule were in effect and what the dog does accords with that rule.

Likewise, for any regularity in action there corresponds *some* rule: a thrush singing its song in the early morning isn't following a rule, even if the song is composed of perfectly regular phrases that the bird repeats with perfect exactness. Mere automatic behaviour is therefore not enough in order to be said to be following a rule. Instead, both meaning and rule-following seem to require some intention on behalf of the rule-follower when following a rule, the speaker needs to be somehow justified, guided or instructed by the rule in his actions for it to count as an instance of rule-following—and likewise for meaning, as a trained parrot might consistently squawk out the word "red" in the presence of a red thing and yet we would not say that the parrot *meant* "red" by its squawk.

We therefore need a slightly stronger formulation of what (R) is intended to capture:

(RG) A term has meaning if and only if its use is *governed* by a rule.

On this version of the principle, it is not enough to merely say or do the correct thing to mean a particular thing by one's utterance, but to do it *because* of the corresponding rule in some appropriate sense—guided by it.

Notice, however, that (RG) is stronger than the claim that whenever there are correctness conditions, there are rules, in a non-trivial way: if we accept (RG) as the correct way to cash out the idea that meaning has correctness conditions, it implies that rules are more fundamental than meaning: it is *because* our use of words is rule-governed that they have meaning, and the rule, and the fact that it is being followed, is what makes this possible—that there is correctness and incorrectness is in some sense due to or determined by the rule. The rule, as it were, prescribes to the agent what to do in order to follow the rule and it is the understanding of what

the rule requires that makes the understanding of the concept that the rule governs possible. It therefore cannot be, on this view, that the correctness conditions of terms are primary and those of rules derive from them, because there would be nothing to provide the correctness conditions for the concepts except some rule, itself needing correctness conditions. The rules are what constitute the meaning of the term they govern—if it were not for the rules, there would be no such meaning.

This view, therefore, holds that to explain how the terms and concepts of our language have correctness conditions, we first and foremost appeal to the existence of rules that govern them and likewise that for someone to learn a language, they need, in some way or another, to grasp or cotton on to the rule and subsequently follow it in their linguistic practices. This does not, of course, mean that the speaker must be able to say or even say to himself how he grasps the rule.[2] The point is rather that grasping and being guided by rules is taken to explain the meaning of words and how agents can understand that meaning, not the other way around. Call this the "primacy of rules thesis".

This principle, I take it, is what underlies the idea that language is governed by rules. It is not that we first learn how to speak, however we might be able to do that, and *then* learn how to follow rules, as a linguistic practice, but we learn how to speak by grasping the rules that govern our language. In the context of Wittgenstein's philosophy, I will sometimes refer to the idea that language is rule-governed, i.e. accepts (RG), as the *calculus view*.[3]

The chapter will be in three parts. In the first two parts, I'm going to present two problems for the idea that language is governed by rules, the first due to Katrin Glüer and Peter Pagin, and the second to Crispin Wright. In the last part, I will argue that contrary to the orthodox reading of his later philosophy, Wittgenstein did not in fact conceive of language as rule-governed, and in fact, the rule-following paradox is one of his main arguments against that view.

2. See e.g. Dummett (Dummett 1986, p. 464): "Whatever the full content of Wittgenstein's distinction between an *Auffassung* (way of grasping) and a *Deutung* (interpretation) may be, it at least means that one who grasps a rule or understands a sentence need not be able to *say* how he understands it. He does not have to be able to say it even to himself."

3. We might also refer to the view that language is rule-governed *semantic generalism*, and the opposite as *semantic particularism* on an analogy with moral particularism, as the view that Wittgenstein was a semantic particularist is fairly common among moral particularists. They see themselves as applying Wittgensteinian arguments about meaning to ethics. See e.g. Dancy 2004.

I will not argue for an alternative view here, as that is the task for the next chapter, but merely to point out the problems that dog us if we accept primacy of rules-thesis and how much simpler our philosophical life would be, if we could find another way to provide correctness conditions for our concepts than that of rules. The idea, then, is to set up a dialectic such that it becomes plausible to look elsewhere for the constitution of meaning. The point is not, however, I should emphasise, that we never actually follow rules, even when speaking a language, but rather that rules do not play an explanatory role in accounting for why we have this ability nor for why words have meaning: we do not learn a language by learning how to follow the rules that govern its use—there are no such general rules—but rather, we learn how to follow rules by learning how to speak a language. In particular, there is such a thing as correct or incorrect use of a word or a concept, but this is not to be explained in general by the grasp of rules.

In other words, semantic generalism has the direction of explanation backwards. It is not that no aspect of language is rule-governed, that is undoubtedly true for many important aspects of it, but rather that rules cannot be the most basic phenomenon in accounting for meaning—it's not rules all the way down, and to understand meaning, we do not need to first understand rules. Furthermore, on the account that I will develop, (R)—the principle that whenever there are correctness conditions, there are rules, is not rejected, merely the stronger principle (RG) that the rules are what *determines* correctness.

## 3.1 Wright's modus ponens model of rule-following

In this section, I'm going to examine an argument by Crispin Wright against the idea that rules can guide agents in basic cases. The argument has some connection to the discussion of truth-conditional semantics in the previous chapter, but crucially, Wright does not conclude from it that rules do not constitute meaning.

Wright's starting point is something like the primacy of rules thesis, i.e. that it is by grasping rules determine correct and incorrect action that we can come to understand language. He writes:

> It is natural to think that in any area of human activity where there is

> a difference between *correct* and *incorrect* practice, which we achieve is (partly) determined by rules which fix what correct practice consists in, and which in some manner guide our aim. (Wright 2007, p. 481)

He also claims that it is a platitude that "wherever there are rules, there have to be *facts* about what their requirements are" (ibid., p. 481) and these facts must be such that we are able to appreciate that they obtain, if we are to be said to receive guidance by the rule.

These very reasonable assumptions lead to a paradox, however. Consider a case where *S*'s use of language is in fact rule-governed, in the sense that there is some rule that *S* is guided by in his linguistic practices and suppose we are thinking of some basic predicate, such as 'red'. In this case, the rule *S* employs and how he does it, would be something like the following (on what Wright calls 'the modus ponens model' of rule-following):

(Rule) If …$x$ …, then it is correct to predicate 'red' of $x$.

(Premise) …$x$ …

(Conclusion) It is correct to apply 'red' to $x$.

As Wright points out, if we want to include cases like this in the modus ponens model, we require an 'anterior concept' which determines whether or not the right conditions obtain for the application of the rule, namely the one indicated by '…$x$ …'. If 'red' really is basic, this concept cannot be anything else than *red* and so the ability of the rule to guide *S*'s actions seems to have evaporated, since if we try to rectify the situation by putting something else into the slot occupied by '…$x$ …' in the rule, that concept, whatever it is, would have to be basic, and we can repeat the preceding reasoning, leading to a regress.

For Wright, this shows that in basic cases rule-following is uninformed by

> anterior reason-giving judgement - just like the attempts of a blind man to navigate in a strange environment (ibid., p. 496)

and that in such cases "we do not really *follow* – are not really guided by – anything" (ibid., p. 496). And so, Wright's response is to go *quietist* and conclude that in basic

cases, there cannot be anything like a substantive account of rule-following, and further that in such cases our application of the rule cannot be explained by "any appreciation about facts about what the rules require" (ibid., p. 498).[4]

The basic dilemma presented by Wright's argument is that we either need to abandon the idea that language is at bottom rule-governed (i.e. constituted by rules that guide agents) or find a different model of rule-following, a less natural one, perhaps, than the modus ponens model.

### 3.1.1 Rule-following and truth-conditional semantics

In the last chapter, we examined an argument that purported to show that if accept a certain substantial conception of fact, there could not be any facts about what a speaker meant by his words. This argument involved an implicit detour from the intermediary conclusion that no mental facts can secure the application of a rule, through the idea that meaning in general involves there being some fact about the speaker, substantially understood, in virtue of which something is meant, to the conclusion that no facts, and thus not truth-conditions, can serve as the basis of meaning.

It might very well be true that Wittgenstein would have endorsed this conclusion, but it is likewise true that he simply does not make this argument. It rather seems that his argument against truth-conditional semantics is made via his considerations of rule-following. One philosopher who has understood Wittgenstein's arguments in this way is Claudine Verheggen (2003). In her paper, Verheggen is mostly concerned with arguing against certain quietist readings of Wittgenstein's remarks, mostly those of McDowell (1984) and Finkelstein (2000). They emphasise Wittgenstein's own expression of the paradox as that of always looking at the grasping of a rule as an interpretation, a view they see having its roots in the idea that there is a gap between a sign and its meaning. They argue that as soon as we realise that this question is never raised in our everyday practice, we will cease to ask it, as we will automatically see that our signs are meaningful and alive, dissolving the paradox. If we reject this thesis that signs stand in need of interpretation to be meaningful, the

4. See also Wright 2012 for discussion.

> problems about the normative reach of meaning, …since they depend on a thesis that we have no reason to accept, stand revealed as an illusory. (McDowell 1992, p. 48–49)

Verheggen disagrees with the quietists both about this conclusion and the root of the paradox. For her, Wittgenstein certainly asked the question of how signs can attain meaning (rather than merely dissolving them) but at the same time, this question is not the most fundamental cause of the rule-following paradox at all. It it arises for Wittgenstein, Verheggen says, precisely because he has given his own answer to this question, namely that the meaning of signs is their use. In the passages on rule-following, Wittgenstein is grappling with the implications of that view, more specifically, how we can account for the objectivity and normativity of our utterances if the meaning of our words is their use: how can the use of a word, which is finite, determine how I should use it in the future and how can it establish that some ways are right and others wrong (Verheggen 2003, p. 289)? This Verheggen calls the *determination problem* and is of course the same as our question of how meaning is constituted.

Wittgenstein's starting-point in trying to solve this problem, Verheggen says, is by examining a particular answer to the problem, namely *semantic platonism*, which she describes as the view that there are standards outside of our linguistic practices to which our applications must conform in order to be correct, standards which come in the form of extra-linguistic abstract entities, either a rule or a meaning, which we grasp when we understand a word. These abstract objects, Verheggen thinks, need to be interpreted in order to do their job, and in this case, anything that 'comes before the mind' is itself 'semantically opaque or indeterminate', leading to a regress of interpretations. The conclusion is, that no abstract object can ever provide words with meaning (ibid., p. 291).

Semantic platonism, however, is only supposed to be one example of a family of views intended to plug the gap between meaning and sign. One could, Verheggen claims, also appeal to the "features of objects and events" (ibid., p. 292-293) in our environment to provide this standard, which of course is what truth-conditional semantics assumes by its claim that the meaning of a statement is given by its truth-conditions: if grasping the meaning of 'red' is to understand the circumstances under which it is true that something is red, that trivially implies a prior grasp of

the circumstances under which that very thing is red.

Consequently, Verheggen claims that the rule-following paradox shows, not only that entities of a certain kind can provide such a standard, but that it is not possible that

> entities of any kind, conceived of in total independence of linguistic practices, could ever perform that task. (ibid., p. 293)

That, she thinks, is the real root of the paradox and why we are prone to think that following a rule is always an interpretation. She continues:

> The very idea of meaning resting solely on the association of words with extra-linguistic entities is incoherent because we in fact need a language to be able to do what is alleged to provide us with one in the first place. (ibid., p. 291)

But, she says, this any such entity needs to be interpreted, and therefore an association between words and extra-linguistic entities cannot take place before their meaning is established—and that includes properties of objects.

This notion, that Wittgenstein's rule-following paradox is a powerful argument against such a position is fairly common. Here is Wilson on the matter:

> [If] a set of mind and language-independent properties were to be established as the standard of correctness for a term 'T', then users would have to have some kind of pre-linguistic "grasp" of properties-in-the-world that allowed them to form the semantic intentions that purportedly establish certain of the properties as the standard in question. (Wilson 1998, pp. 109–110)

Here, a term 'T' is for Wilson something that cannot be reduced and explained by other terms, the final word to be explained by an ostensive definition. Now, as Verheggen points out, Wilson does not explicitly say why Wittgenstein is supposed to have reached this conclusion, but presumably it is for the same reason she outlined: the paradox shows that extra-linguistic entities, independent of linguistic practices, cannot provide a ground for meaning.

*Prima facie*, this does seem to me to be a good way of thinking about Wittgenstein's own position and what it entails. For instance, much of the discussion in the early part of the *Philosophical Investigations*—discussion that very clearly prefigures the rule-following considerations, e.g. §§28–32—seems to be aimed at undermining this very picture: either $A$ grasps what a property $\varphi$ is by an explanation or example, if the former, then we can repeat the question of how $A$ grasps the properties used in that explanation, until we reach a final explanation which is by example only. But then any term 'T' which is supposed to apply to some basic property is always ambiguous and empty of content, since it can always be interpreted in various ways: we can never get a grip on any explanation without assuming some kind of background practice, and this picture cannot provide it.

And this of course is the problem that beset the modus-ponens model: it seems to construe language, as Wright himself points out (Wright 2007, p. 495), in terms of the Augustinian picture of language that Wittgenstein describes in the beginning of the *Investigations*, not as a confusion between meaning and naming, but by presupposing that we already have a grasp of the concept a term refers to, the very term whose meaning the rule is supposed to constitute. By the end of those series of remarks, Wittgenstein writes:

> Augustine describes the learning of human language as if the child came into a strange country and did not understand the language of the country; that is, as if it already had a language, only not this one. Or again: as if the child could already think, only not yet speak. And "think" would here mean something like "talk to oneself". (*PI*, §32)

That is to say, on the modus ponens model a rule sets up the meaning of a term by associating some particular object with a general concept, e.g. a dog with the concept *dog*, and that seems to require that it is already established that the object in question falls under the very concept constituted by the rule. It is as if it is determined prior to any language or linguistic practice what counts as *cat*, *king*, *four*, etc. and that learning a language is a matter correlating that already existing meaning of the word to the word itself on the one hand, and correlating the word with the object on the other, like attaching labels to a thing[5]. Wright concludes:

5. Which is of course reminiscent of Quine 1968, p. 186.

> In short, the problem with extending the modus ponens model to cover all cases of rule-following, including that involved in basic cases, is that it calls for a conceptual repetoire *anterior* to an understanding of any particular rule – the conceptual repertoire needed to grasp the input conditions... (ibid., p. 496)

McDowell makes a similar point as follows:

> Wittgenstein's reflections on rule-following attack a certain familiar picture of facts and truth, which I shall formulate like this. A genuine fact must be a matter of the way things are in themselves, utterly independently of us. So a genuinely true judgement must be, at least potentially, an exercise of pure thought; if human nature is necessarily implicated in the very formation of the judgement, that precludes our thinking of the corresponding fact as properly independent of us, and hence as a proper fact at all. (McDowell 1984, p. 351)

One might object that this merely shows that the modus ponens model is too simple to capture what is really going on. That might very well be true, but it is hard to see what it might be to follow a rule if there cannot be any prior facts as to what that rule requires, and for the agent to grasp those facts, they already need to know what facts of the sort that are constituted by rules, leading to a regress. This would presumably be the case on *any* reasonable conception of what it is to follow a rule, even if we could do better than the modus ponens model.

In the next section, I will take a look at a different argument agains the notion that language is governed by rules, due to Glüer and Pagin. In their paper, they examine a few different ways the doctrines that meaning is constituted by rules and rules guide agents in their linguistic practice are incompatible, and so survey a few different ways of spelling this out, other than the modus ponens model.

## 3.2 Glüer and Pagin on constitutive rules and practical reasoning

Searle's distinction between *regulative rules* and *constitutive rules* is a familiar one (Searle 1969, pp. 33–42). According to this distinction, regulative rules regulate some activity whose existence is logically prior to the rules that regulate it and constitutive rules *constitute* an activity whose existence depends on and cannot be explained or described without reference to those very same rules. They often have the form "Do **Φ**!" or "When in circumstances *C*, **Φ**!". A paradigm example of such rules are the rules of traffic. Traffic can and has existed without any rules and explaining what traffic is does not require reference to any rules. The rules that we do have merely prescribe certain behaviours and proscribe others.

The paradigm example of constitutive rules, on the other hand, are the rules of chess. The game of chess logically depends on its rules to be the game that it is and no adequate description can be given of the game that does not mention the rules in some way. Not only does chess require rules, it requires *these* particular rules, the rules of chess. Such rules tell us what something *is* and not necessarily what to do.

Thus for Searle, constitutive rules can be said to "create and define new forms of behaviour", as he puts it (ibid., p. 33), since without the rules of chess there could not be such a thing as playing chess, nor intend to do certain things within the game, e.g. mate your opponent or fork their rook and king with your knight, as without the rules, there would be no such thing as intending to fork or mate. Constitutive rules *constitute* some practice—i.e. make that very practice possible, and so make the intention of engaging in that practice possible as well, including the intention of performing a specific action within that practice.

It is natural to think that if language is rule-governed, then it is governed by constitutive rules—rules that constitute meaningful utterances. The rough idea is that the correctness and incorrectness of particular utterances—i.e. that a principle like (C) holds for language—can be explained by the presence of rules, whereby an utterance is correct if and only if it is performed in accordance with the rule. This idea, it might plausibly be held, can also explain how agents can learn and use language: it is by *grasping* the rule that we learn the meanings of words and

by subsequently following them, we can use those words correctly. The rule, we might say, is then the *reason* for the agent to say one thing rather than another when using language.

In this section, I want to take a look at arguments due to Glüer and Pagin (Glüer and Pagin 1999) that aim to show that the notion that a rule can both constitute meaning and the idea that rules guide an agent in their linguistic use cannot be so easily combined. They contrast two ways of understanding what constitutive rules are, and further argue that neither conception can do the dual job the proponent of the idea that language is rule-governed wants them to do.

**The traditional reading** On the traditional reading of constitutive rules, which Glüer and Pagin attribute to Searle (1969) and G. C. J. Midgley (1959), constitutive rules determine a new type of action $\Phi$ in a context $C$ by connecting or identifying it with some action type $\theta$ which can be described without the means of the rule.[6] The standard formulation of such a constitutive rule for Searle is something like the following:

**(CR)** Doing $\theta$ in context $C$ counts as doing $\Phi$.

Accordingly, meaning-constituting rules would look something like the following example from Glüer and Pagin (Glüer and Pagin 1999):[7]

**(R1)** Uttering 'green' counts as expressing the concept *green*.

Rules of this form are generally, as they put it, "doubly constitutive". They both constitute a whole practice, e.g. a game, by being a part of its system of constitutive rules, and constitute a certain act within that practice. For example, the rule

**(G)** Passing the whole ball over the goal line between the goalposts and under the crossbar while playing football counts as scoring a goal

is both a part of the system of rules that constitute football and constitutes the action of scoring a goal by connecting the action of passing of the ball over the goal line with the new action type of scoring a goal.

6. For discussion, see also Reiland 2019.
7. Perhaps we should add '...in an English speaking context'.

In contrast, prescriptive rules only mention one kind of action type, for example, 'Score a goal!' or 'When you have the goal open in front of you, score!'. However, as Searle himself points out, there is a problem. Any prescriptive rule can be brought into the form of a constitutive one. For example, the prescriptive rule "Officers must wear ties at dinner" can easily be converted to the constitutive-looking rule "Non-wearing of ties at dinner counts as wrong officer behaviour" (Searle 1969, p. 36) where the non-wearing of ties at dinner is seemingly made into the new action type 'wrong officer behaviour'.

Be that as it may, we can still ask, with Glüer and Pagin, how language can be said to be rule-governed on what they call the Midgley-Searle model (Glüer and Pagin 1999, p. 217)? Recall that this notion essentially involves the agent being *guided* by the rule when speaking a language. Glüer and Pagin, reasonably enough, understand this idea of guidance by a rule as the agent having the rule as a *reason* for his acting in the way that she did—in the case of language, somehow guiding the speaker in making her utterances. Glüer and Pagin are not concerned with whether those reasons are good reasons nor with other decision-theoretic elements of the agent's action—what matters is whether or not the reason actually motivated the agent in his action or is a candidates for being such a reason. Glüer and Pagin call such a reason a *basic reason* (ibid., p. 209).

Consider first what Glüer and Pagin call the *simple inference*—an intuitive conception of how a rule might guide our action (and quite reminiscent of the modus ponens model):

(R) When in $C$, $\Phi$

(B) I am in $C$

(I) So, I shall $\Phi$

Here, (R) refers to a rule that might serve as such a basic reason, (B) some belief the agent has, in this case that the antecedent of the rule actually obtains, and (I) the intention of the agent formed by his practical reasoning.

The first thing to notice is that this schema requires a prescriptive rule, not a constitutive one. There is simply no slot for the constitutive rule, and hence it

cannot guide the agent. We might try to shore this up by changing the schema like this (ibid., p. 217):

(PA) I want to do what *R* requires

(B) *R* requires that I **Φ**

(I) So, I shall **Φ**

Here, (PA) is an attitude towards the rule by the agent (called a 'pro-attitude' by Glüer and Pagin). This looks more promising, but since it is not obvious by the form of the schema that there is no slot for the constitutive rule. However, if the rule *R* is a constitutive rule on the Midgley-Searle model—i.e. of the form 'Doing $\theta$ in context *C* counts as doing **Φ**'—there is simply nothing that it requires and no way to not **Φ** while $\theta$-ing in *C*, and thus no way to violate the rule (ibid., p. 217). No agent is *required* to anything if a constitutive rule of the Midgley-Searle form is in force and there is no way for the agent *not* to do **Φ** by doing $\theta$ in the circumstances provided by the rule: a football player for whom (G) is in force is not required to pass the ball over the goal line and cannot avoid scoring a goal by doing so. Glüer and Pagin's challenge to this model is, if it is not possible for *S* to violate the rule, how could it possible be guiding *S*?

We might try a different model, more appropriate to speech acts. Consider for instance the following pattern where some collection of rules like (R1) is in force for me:

(PA) I want to say that *p*

(B) Making the utterance *s* counts as saying that *p*

(I) So, I shall make the utterance *s*

Here, it might seem that the rule is in fact guiding me in my speech acts. Glüer and Pagin, however, point out that the rule does not have motivational role for the agent in this pattern, but merely a doxastic one (ibid., p. 218). We might just as easily have slotted in a prescriptive rule there, for instance, whereby the agent had accepted the rule "If you want to say *p*, utter *s*". It is in fact completely inessential

that there is a rule at all—all we need is that the agent believes that there is some fact about meaning, and that need not be provided by a rule on this model.

Glüer and Pagin further argue that if we'd maintain that it was *possible* for a Midgley-Searle rule to constitute meaning by being a doxastic reason for the agent, we'd lose the notion of it constituting meaning since the rule would then merely be a way for me to choose a different manner of expressing my thoughts, which for Glüer and Pagin, implies that they are constituted by something other than the rule, since would be only by accepting the constitutive rule in the first place that I'm able to express that thought at all, and given that expression, I would not be able to use that expression to express any *other* thought (Glüer and Pagin 1999, p. 219). It therefore seems that a constitutive rule of the Midgley-Searle form could either constitute the meaning of an utterance, or guide an agent in his speech acts, but not both. This is of course reminiscent of the argument rehearsed in the last section.

**The non-traditional reading** Perhaps, however, the problem lies with the Midgley-Searle conception of constitutive rules, not the intuitive idea itself, namely that certain activities such as football, chess—or indeed language—could not be engaged in unless there are some rules in force. This does not require, as Glüer and Pagin point out, that the constitutive rules have the form of a Midgley-Searle type rule.

Instead, they consider a suggestion by John Rawls (1955) to the effect that action types belonging to a practice constituted by rules are actions that cannot be performed outside of the 'stage-setting' of the practice (1955, p. 27). This stage-setting cannot be anything physical or physically specifiable, Glüer and Pagin think, since

> 22 men may run around kicking the ball exactly as if there were a game, but if the game isn't on, still no goals are scored. (Glüer and Pagin 1999, p. 221)

This, they think, indicates what is "essential" to the idea of stage-setting: that that the rules are in force for the agents. When we play a game of football, we've decided that the rules are to apply to us, we decide what to count as the field, what

a goal, and so on, and as we play, what we do is evaluated by the rules set up by us. Practice, then, is constituted by a set of rules if it is possible to engage in that practice "only insofar as the rules of that set are in force for the agent" (ibid., p. 221). It then follows that it is only if the rule (G) is in force that I can score a goal by passing the ball over the goal line, otherwise my action was just that, passing a ball over a painted line.

How might we try to apply *this* notion of constitutive rules to the idea that language is rule-governed? The rule (R1) is no good, since that was of Searle-Midgley form, but we might try the following rule (ibid., p. 222):

**(R2)** 'green' is correctly applied to an object $x$ if and only if $x$ is green

This seems promising, and in line with our principles (C) and (C′). This rule, or another of that form, could enter into the 'motivational slot', as Glüer and Pagin put it, of some piece of practical reasoning (ibid., p. 222). Consider for instance the following argument schema:

(PA) If an utterance of $s$ is correct, then I want to make the utterance $s$

(R) Uttering $s$ is correct if and only if $p$

(B) $p$

(B) So, an utterance of $s$ is correct

(I) So, I shall make the utterance $s$

Glüer and Pagin identify two problems with this schema. The first problem is that even if the rule is in force for an agent, that does not necessarily mean that the agent has the 'pro-attitude' towards the action that they consider correct. For instance, in football it is not allowed to touch the ball with the hand, and yet it is not necessary for the agent to have a pro-attitude towards the corresponding rule in order to play football. All that is needed is that the agent *accepts* the rule and that is completely consistent with having a pro-attitude *towards* touching the ball with the hand, and thus a pro-attitude for the opposite rule. In the case of language, I could likewise accept that a rule is in force for me, and yet go against it, for example in the case of irony, and hence having that pro-attitude towards the

rule is not necessary for a meaningful utterance, even if it is necessary me to take it to be in force for me to be able to make that utterance. The rule, therefore, does not have a guiding role on this model. (Glüer and Pagin 1999, p. 223).

The second problem is more decisive in their estimation, however. The schema just considered is not apt for the role the rule needs to play, because usually we do not just merely want to say what is correct, but rather to make a particular speech act, e.g. asking whether that *p*, asserting that *p* and so on. Naturally, we generally want our utterances to be correct, but usually that is not our aim in itself (we don't sit around making correct utterances outside of any further context, for instance), and here the rule is useless: if I want to say that *p*, how does a rule that tells me that *s* is correct if and only if *p*, help?

The result of putting *that* rule, Glüer and Pagin point out, instead of a Searle-Midgley rule into the relevant place into some piece of practical reasoning is simply incoherent (ibid., p. 224). In the following argument, for instance

(PA) I want to say that *p*

(R) Uttering *s* is correct if and only if *p*

(PA) So, I want to utter *s*

the conclusion simply does not follow from the premises as the rule isn't connected to the other premise and the conclusion in the right way.

If we'd try again to shore this up by pointing out that we might infer from the premise that (R) is in force that *s* does mean *p*, and hence that uttering *s* would count as saying *p*, Glüer and Pagin counter that if that were so, the rule would not be playing a guiding role in that inference, but merely a constitutive one. Any meaning-determining fact, however that is specified, could play that role. Furthermore, it is not even necessary for the agents, in general, to know about such facts to be able to speak a language at all, even if this particular inference might be salvageable (ibid., p. 224).

Glüer and Pagin conclude from their discussion that the insofar as some plausible assumptions are made about practical reasoning, that the combination of views required by the calculus view, that rules can both be constitutive of meaning and provide guidance to agents in their linguistic utterances, is incoherent. Some rules

can determine meaning, but not generally guide meaning, e.g. those just discussed, but others, like Searle-Midgley rules or instrumental rules can guide agents, but only in a capacity as a fact about meaning, and not themselves constitute it.

Glüer and Pagin do not give an *a priori* argument that language cannot be rule-governed—in the sense that agents are guided by rules in their use of it, but rather survey a few plausible ways to cash that idea out and find them wanting. However, as they point out, the idea that meaning is determined by constitutive rules in the first place is very often motivated by the notion that we could thereby account for how agents gain their competence with language, and if that turns out to be more problematic than it seems, our motivation for holding that view seems not as strong.

In the next chapter, I will argue that meaning is not in fact determined by constitutive rules, but what I will call *constitutive practices*. First, however, I want to examine the exegetical question of whether Wittgenstein conceived of language as rule-governed as an exegetical matter.

## 3.3 Does Wittgenstein think language is rule-governed?

A common way to read Wittgenstein's overall philosophy of language, and especially his remarks on rule-following, is to attribute to him the view that language is fundamentally a rule-governed activity: speaking a language requires adherence to or guidance by rules. This reading is so common that we might call it the received view about the import of Wittgenstein's remarks on rule-following.[8]

According to this reading, while Wittgenstein does indeed hold some kind of use-theory of meaning, this aspect of his thought is mediated through the intermediary step that rules are what governs the use of words. In particular, the meaning of linguistic expressions necessarily involves linguistic agents following or being guided by rules that govern the use of their expressions. Accordingly, Wittgenstein's point with the rule-following remarks is not to reject the view that language is rule-governed, but rather to save it from a particularly threatening paradox. He

8. See for instance Kripke 1982; Hilmy 1987; Glock 1996; Hacker and Baker 2009. Hacker and Baker's view is a bit ambiguous. In an earlier work, they make a distinction between 'humdrum' rules of language that we teach children and a more theoretical conception they find misconceived (Hacker and Baker 1984b). In general, they seem happy to play both sides of the fence, however.

is concerned to show that we are liable to be misled into profound philosophical difficulties by thinking of how we follow or are guided by rules in certain natural, but ultimately confused ways, but when those misunderstandings have been cleared away, rules remain to be of the most fundamental importance to a correct account of linguistic meaning. As before, I will refer to this view—that linguistic meaning is essentially rule-governed—as (RG) and a reading that attributes this reading to Wittgenstein's later work, the 'received view'.

In this section I argue that it is in fact a mistake to read Wittgenstein as endorsing any form of (RG). By this I simply mean that he rejects the idea that meaning is essentially a matter of following rules that govern the use of words and that in order to understand the meaning of a word a linguistic agent need not internalise any rules at all (and in fact, that the upshot of Wittgenstein's rule-following paradox is indeed that this is impossible).

The point is *not*, I should emphasise again, that we never actually follow rules, even when speaking a language, but rather that rules do not play an explanatory role in accounting for why we have this ability or for why words have meaning: we do not learn a language by learning how to follow the rules that govern its use—there are no such general rules—but rather, we learn how to follow rules by learning how to speak a language and taking part in the surrounding practices. For Wittgenstein, following a rule is a particular kind of language-game, not something that lies behind all such games. In other words, (RG) has the direction of explanation backwards. It is not that no aspect of language is rule-governed, which is undoubtedly true for many important aspects of it, but rather that rules cannot be the most basic phenomenon in accounting for meaning—it's not rules all the way down.[9]

9. This section is heavily influenced by Kathrin Glüer and Åsa Wikforss's paper "Es braucht die Regel nicht: Wittgenstein on Rules and Meaning" and I will cite it heavily throughout. My arguments for this position sometimes deviate from them, however, and I prefer to state the conclusion is stronger terms, as they nowhere attribute semantic particularism to Wittgenstein. I suspect, however, that this is mostly a question of emphasis.

Recently, Severin Schroeder (2017b) has presented arguments for a similar conclusion.

### 3.3.1 The analogy with games

It is quite common to divide Wittgenstein's later philosophical development into two distinct phases, a 'middle phase' and the 'late'. In the middle phase, Wittgenstein roughly held the view that language was a collection of independent calculi where the meaning of words is defined by the rules of the calculus, while in the latter, his emphasis moved to language-games rather than calculi.[10] On the received view, Wittgenstein never fully abandoned his former view but merely refined it: language-games are governed by rules and when speaking a language, the speaker's speech acts still instantiate some rule. He therefore always remained committed to a version of (RG). The following passage, quoted by Glüer and Wikforss in their paper on the matter, is instructive:

> Wittgenstein did not abandon the idea that language is rule-governed, he clarified it, comparing language no longer to a calculus but to a game. Unlike these analogies, the idea that language is rule-governed is not just a heuristic device. Understanding a language involves mastery of techniques concerning the application of rules. (Glock 1996)

Glock's last claim is especially important in understanding the contrast between (RG) and its negation, semantic particularism: For Glock, learning how to speak a language is a matter of mastering the application of rules. For the semantic particularist, mastering the application of rules is a matter of learning how to speak a language.

Glüer and Wikforss point out that Glock is in a certain aspect correct, Wittgenstein does make an analogy between language and games, but at the same time that he takes this analogy "all the way". For them, Wittgenstein's later remarks on rules are directed *against* (RG), and not intended as a subtle defense of it, and while Wittgenstein's discussion is often framed in terms that make it 'tempting to take him literally', the analogy, as analogies are wont to be, is fundamentally limited. This analogy is exactly just that, an analogy (Glüer and Wikforss 2010, p. 3-4).

Some of Wittgenstein's corrections to the *Big Typescript* lend credence to this reading of the analogy as being limited. In a chapter having the very descriptive

10. See e.g Hintikka 1989, Gerrard 1991 and Shanker 2005.

heading "Language Functions as Language only by Virtue of the Rules We Follow in Using it, just as a Game is a Game only by Virtue of its Rules" Wittgenstein wrote the following corrective remark:

> That is not correct, in so far as no rules have to have been laid down for language; no more than for a game. But one can look at language (and a game) from the standpoint of a process that uses rules. (*BT*, 196)

Here, what is being corrected is, as the title of the chapter indicates, is precisely the view that language is what it is in virtue of rules governing its use. In another correction in the same chapter, Wittgenstein wrote:

> "Contrat sociale" – here too, no actual contract was ever concluded; but the situation is more or less similar, analogous, to the one we'd be in, if…And there's much to be gained in viewing it in terms of such a contract. (*BT*, 196)

These remarks, explicit corrections to claims that language is rule-governed, seem to be saying that viewing language as a rule-governed game is merely an analogy and that it can merely be useful to regard language in this way.

There seems therefore to be something right about this idea that Wittgenstein's analogy between games and languages should not commit us to language being rule-governed, as Glock seems to think. First of all, there is the obvious point that Wittgenstein's own discussion of games is meant to bring out the fact that 'game' is a family-resemblance concept (and indeed the paradigmatic example of such a concept). Such a concept, Wittgenstein thinks, is precisely a concept which is not 'everywhere circumscribed by rules' (*PI*, §68).

That's not to say that games might not often be rule-governed, undoubtedly they are, but there is nothing in Wittgenstein's discussion of games that indicates that he thought of games as fundamentally rule-governed activities, and therefore it is hard to see how comparing language to games is meant to clarify the notion that language is fundamentally rule-governed. The rejection of (RG) is not meant to signal that using language is *never* a question of following rules—that might be the case sometimes—but that rules are not the more basic phenomenon through

which we can then *explain* language, just as rules are not the way we explain games in general, although *some* games, even most, might best be explained in this way.

Furthermore, Wittgenstein sometimes uses games as an example to argue for the *opposite* view. The first example I will discuss is in §§81–84 of *PI*. In §81, Wittgenstein writes:

> F. P. Ramsey once emphasized in conversation with me that logic was a 'normative science'. I do not know exactly what he had in mind, bit doubtless closely related to what only dawned on me later: namely, that in philosophy we often *compare* the use of words with games and calculi which have fixed rules, but cannot say that someone who is using language *must* be playing such a game.—But if you say that our languages only *approximate* to such calculi you are standing on the very brink of misunderstanding.[11] (*PI*, §81)

And a little later in the same remark:

> All this, however, can only appear in the right light when one has attained greater clarity about the concepts of understanding, meaning and thinking. For it will then also become clear what may lead us (and did lead me) to think that if anyone utters a sentence and *means* or *understands* it he is operating a calculus according to definite rules. (PI, §81)

The discussion that Wittgenstein refers to, on what it is to understand, mean and think, are most likely the passages on rule-following. If so, Wittgenstein explicitly says that his arguments in this regard are aimed at the idea, which he here says he held previously, that uttering a sentence is the same as 'operating a calculus according to definite rules', which is nothing but (RG). In my view, this remark should on its own provide pretty strong evidence for the reading that Wittgenstein rejects (RG), and in particular, that is at least one of his targets with the discussion on rule-following. If Glock's reading is right, and Wittgenstein *does* think that

11. Here it might be significant that the German original is slightly different. Wittgenstein writes: "wir nämlich in der Philosophie den Gebrauch der Wörter oft mit Spielen, Kalkülen nach festen Regeln, *vergleichen*". 'nach festen Regeln' does not seem to modify both 'Spielen' and 'Kalkülen' as in the English translation, because of the comma in the original.

language is rule-governed, it is hard to see what view Wittgenstein is rejecting here.

In the next remark (§82), Wittgenstein brings up the objection to (RG) that it is unclear what an expression such as "the rule by which he proceeds" means when we considers someone's use of words (since questioning and observation always underdetermines the rule in play). He then writes:

> Doesn't the analogy between language and games throw light here? We can easily imagine people amusing themselves in a field by playing with a ball so as to start various existing games, but playing many without finishing them and in between throwing the ball aimlessly into the air, chasing one another with the ball and bombarding one another for a joke and so on. And now someone says: The whole time they are playing a ball-game and following definite rules at every throw.
>
> And is there not the case where we play and–make up the rules as we go along? And there is even on where we alter them—as we go along. (*PI*, §83)

It seems that if Wittgenstein wants to draw an analogy between language and games, this analogy is *not* intended to show that language is fundamentally rule-governed, since what he describes here is exactly supposed to show that the players are *not* playing with definite rules in mind when they go about their game.

In the very next remark, Wittgenstein says: "I said that the application of a word is not everywhere bounded by rules. But what does a game look like that is everywhere bounded by rules?" (*PI*, §84). This of course recalls Wittgenstein's earlier discussion in §68 where he first introduced the idea of family-resemblance concepts, which for him are concepts which are not rule-governed at all. There he compared them to tennis, which is a game 'not everywhere circumscribed by rules' but "a game for all that" (§68). Again: if Wittgenstein's point with the analogy to games is to say that language is like a game, he does not seem to be emphasising their rule-governedness, but rather the fact that games are typically *not* rule-governed in every aspect.

There is a further reason here to reject this reading. It is an overarching theme of Wittgenstein's later philosophy, going all the way to the *Blue Book* at least, to reject any general explanation of particular linguistic phenomena, in the sense that there is something that they all have in common. For Wittgenstein, it is simply a fallacy to say that because we use the same word in different cases, then there must be something which the two cases have in common. That is the underlying thought with his discussion of family-resemblance concepts, that they cannot be circumscribed in advance. If that is the case, there is something odd about saying that while two things do not have to have anything specific in common in order to fall under the same concept (i.e. accepting that there are family-resemblance concepts), at the same time, our use of language (and concepts) is nonetheless rule-governed. That is to say, it makes very little sense to say that it is a rule which governs the concept 'game' that *x* and *y* are both games and at the same time maintain that nothing general determines in advance that *x* and *y* both fall under the same concept. Isn't that what a rule on this conception is, something general that circumscribes the use of a concept? If Wittgenstein did in fact hold (RG), there would be a inexplicable tension in his view. If one accepts that there are such things as family-resemblance concepts, concepts where two things might not have anything in particular in common, other than falling under that very same concept, then it is hard to see how one could *also* accept (RG). I will return to this line of thought below.

What light is this analogy with games supposed to throw on language then? Here, I believe Severin Schroeder's analysis is correct (Schroeder 2017b). For him, the analogy is mostly supposed to press on us the fact that language is normative—that there are standards of correctness in play. But this, as Wittgenstein's examples seem to suggest, is presumably also the case in games that are *not* strictly rule-governed. On this view, there's nothing inconceivable in imagining, for instance, children playing a game where not only there are no rules laid down in advance, but no clear rules can be found at all, and yet the game has normative elements to it where the children correct and encourage each other—and would be right to do so. Schroeder's description of how this is supposed to work is that the kind of normativity our use of concepts has is that it is "piecemeal" where the "explanations that we can give manifesting our linguistic competence is always just provisional"

(Schroeder 2017b).

The question still remains, however, how this can be made to work, and Schroeder says little about that. Notice, however, that if we agree that the source of normativity in games can be something else than the rules that supposedly govern them and that the normativity of language has the same (or a similar) source, the rule-following paradox is dissolved: rules would inherit their correctness conditions from language, and not the other way around. Of course, this is not the same as actually giving such an explanation for the normativity of language, the point is rather that the problems we are faced with by supposing that (RG) is true are perhaps better solved by bypassing (RG) completely.

### 3.3.2 In the middle period and the *Blue Book*

For Glüer and Wikforss, Wittgenstein still held the view that speaking a language is a matter of being guided by rules when he dictated the *Blue Book*. They point out that there, Wittgenstein makes a distinction between a 'process being in accordance with a rule' and a 'process involving a rule' (*BB*, p. 13). We can only make a distinction between correct or incorrect practice if the expression of the rule form a part of our understanding and meaning. Accordingly, there are, for Wittgenstein, two types or aspects to teaching, one which is a mere drill that sets up a psychological mechanism, such as making the student associate the word 'yellow' with yellow things, and another which provides the student with "a *rule* which is *involved* in the processes of understanding and obeying" (Glüer and Wikforss 2010, p. 9). It is the latter which is of interest:

> Teaching, as the hypothetical history of our subsequent actions (understanding, obeying, estimating a length, etc.) drops out of our considerations. The rule which has been taught an is subsequently applied interests us only as far as it is involved in the application. A rule, as far as it interests us, does not act at a distance. (BB, p. 14)

Those who hold the received view, Glüer and Wikforss say, often cite this passage as an argument for their reading. This, they say, is an "odd exegetical strategy" since it is by and large acknowledged that while the *Blue Book* contains many elements of

Wittgenstein's later philosophy, there are also quite a lot of differences between his views there and in the *PI* (ibid., p. 10). Therefore, whatever we find in there cannot in general be used as an argument for a later view—the latter has to stand on its own. And, I might add, it is hard too see how Wittgenstein of the *Investigations* could have believed that teaching could ever supply the student with a rule in such a direct way, as one of the lessons of the rule-following paradox is that any explanation of a rule necessarily underdetermines the rule.

And in fact, there is a tension in Wittgenstein's views in the *Blue Book* in this regard as well, reminiscent of the one discussed at the end of the last section. Glüer and Wikforss are right to acknowledge that Wittgenstein does not seem to have completely rejected something like (RG) at the time of its dictation. However, there is also quite a lot of resistance to (RG) to be found there, some quite explicit. Consider for instance the following passage which recalls Wittgenstein's discussion about games in *PI*:

> For remember that in general we don't use use language according to strict rules – it hasn't been taught to us by means of strict rules, either. *We*, in our discussions on the other hand, constantly compare language with a calculus proceeding to according to exact rules. This is a very one-sided way of looking at language. In practice we very rarely use language as such a calculus. (*BB*, p. 25)

And a little bit later:

> We are unable to clearly circumscribe the concepts we use; not because we don't know their real definition, but because there is no real 'definition' to them. To suppose that there must be would be like supposing whenever children play with a ball they play a game according to strict rules. (*BB*, p. 25)

Here, Wittgenstein clearly identifies a concept being 'clearly circumscribed' with 'proceeding according to strict rule'. His counter-example is exactly that of a non-rule governed game. For this reason, I think that Glüer and Wikforss are too quick to concede the *Blue Book* to the proponents of the received view—a natural way of interpret Wittgenstein's philosophical development at this point would be to say

that the *Blue Book* is already a part of a transition from (RG) to the rejection of (RG) and therefore a work which is in internal tension in this respect.[12]

A defender of the (RG) reading might explain away these passages by saying that the emphasis should be on 'strict'—linguistic concepts are rule-governed, except the rules that govern them are not strict, where 'strict' would mean something like "allowing some cases to be settled outside of the rule", much like a rule-governed game would be like: a strict rule settles all cases independently of our judgement, a lax one allows some leeway in judgement. There is something odd about that view however. In what sense does a word governed by such a 'soft' set of rules not have an implicit, albeit fuzzy, general definition we have internalised by knowing it? If it does, that seems to go against Wittgenstein's point that words do *not* in general have such a definition, and thus attributing that view to him would be strange, and if it does not, then how should we think about these rules? For if we suppose that on this account of linguistic rules, there would be cases where the rule fully determines our proper use of the concept and cases where the rules do not fully determine our proper use of the concept (which the existence of such a fuzzy rule would seemingly entail), but yet there *is* such a thing as proper or improper use, then there would be a normative-respecting way for us to operate with concepts without the presence of rules governing that concept—there would be cases where we are competent users of the concept without being guided by a rule, at least to some extent. What work are such fuzzy rules supposed to do in a rule-based theory of meaning then? Aren't they simply superfluous? This point of course applies equally to the view if it was Wittgenstein's in *PI*.[13]

For Glüer and Wikforss, Wittgenstein had already rejected this view in the *Brown Book*. There Wittgenstein speaks of cases where the expression 'a game is played according to the rule so and so' is correctly used to describe cases where "where the rule is neither an instrument of the training nor of the practice of the game" but rather as a description of the game (*BB*, pp. 97–98). He then argues

12. Rush Rhees sees the matter much like this in his introduction to the *Blue and Brown Books* (see Rhees 1958, p. x-xii). Hintikka (1989) likewise sees the passages immediately preceding the one quoted by Glüer and Wikforss as an expression of a tension in Wittgenstein's view, not properly resolved until he rejects (RG).

13. There are two other passages in the *Blue Book* which mention language as a calculus, on pp. 42 and 65. Both have to do with the connection between such a calculus and mental states.

that while the expression of a rule might be used in someone's training, it is neither necessary nor sufficient. In the former case we might imagine the student spontaneously use the word properly without any training and in the latter he might interpret the expression of the rule in various, incompatible ways. For Glüer and Wikforss, this marks a rejection of Wittgenstein's earlier point in the *Blue Book*—and that certainly seems right: if Wittgenstein endorsed (RG) for the reasons outlined in the above passage, cited by the proponents of the received view, then he certainly seems to be rejecting *those* reasons in the *Brown Book*. It therefore follows that the passage in the *Blue Book* is not a good argument for his more mature view in *PI*.

### 3.3.3 The rule-following paradox without rules

If we accept that Wittgenstein had rejected (RG) by the time he wrote *PI*, how should we then think about the relevant passages about rule-following, in particular §201 and §202? Recall that in the first, Wittgenstein had claimed that the way out of the paradox was that "there is a way of grasping a rule which is *not* an *interpretation*" and in the latter that "obeying a rule is a practice"). For Glüer and Wikforss, the regress problem from the middle period is what is concerning Wittgenstein in these passages. Whereas there he had sought to solve the problem by appealing to something "that gives a 'final interpretation', he now rejects the suggestion that meaning is determined by interpretations in the first place" (Glüer and Wikforss 2010, p. 11). They continue:

> For a rule to guide the speaker, Wittgenstein holds, an expression of the rule has to be involved with the speaker's use of terms. However, any expression can be variously interpreted; consequently, the idea that meaning is determined by rules leads to a regress of interpretations. (ibid., p. 11)

It's not clear, however, from their exposition, why 'an expression of the rule' has to be involved with the speaker's use of the term. After all, someone committed to (RG) might say, for instance, that the rule-follower simply follows the rule implicitly in some way, without any expression of a rule ever being involved, and

thus that no interpretation is ever needed, or perhaps that rule-following consists in something completely different.[14]

This manoeuvre seems to land us in a dilemma, however: if rule-following is supposed to be a rational activity, then I would have to be able to reason about what the rule requires me to do, especially in a new case I have never encountered before. In this case, I would need to appeal to the rule in my deliberations or judgement about the rule, and this involves an appeal to some expression of the rule and the particular situation in which I find myself (or so I take the idea of 'rational rule-following' to consist in: being able to say or think about what the rule requires and how those requirements obtain). But then, I will either have to interpret the rule and what it requires of me, setting me off on the regress, or abandon the idea of rational rule-following wholesale—and follow the rule 'blindly'.[15]

If we reject the calculus view of language (RG), however, the regress (and the dilemma) are avoided: learning how to understand locutions such as "+2" is fundamentally not about learning a rule at all, and thus there is no regress. That might in turn seem paradoxical, since it seems tantamount to saying that we never follow rules, not even in cases of explicit rule-following. That would be to miss the point, however, which is not that we do not follow rules, but rather, again, about the order of explanation. To explain what it is to mean something by a term by appealing to rules cannot work, because some terms refer to rules and will thus involve an appeal to the notion of a rule, the very thing to be explained.

Similarly, a person's competence in judging that 1002 follows 1000 when following the order "add two" should thus not be explained by him having internalised or 'grasped' a rule, but rather being able to competently judge what the order *means*—and that means something like: knowing how to play the language-game of following this kind of order. On this picture, that is what following a rule *is* and that explanation does not make reference to a notion of a rule. That should not be considered paradoxical at all, though admittedly, much more needs to be said before this can be accepted as an explanation—on the contrary, it is the opposite view which is riddled with paradoxes. On the view presented here, there

14. I also think that that is ultimately Wittgenstein's point: the expression of the rule *does not* have to be involved.

15. This, I believe, is the motivation behind Wright's conclusion that we do in fact follow a rule blindly.

is no regress, since knowing the meaning of a word is having the ability to use it in a whole language-game, and this includes words that purport to refer to rules, and thus there is further reference to rules standing behind the expression of the rule.

Hintikka (1989) sees the matter in quite the same way. For him, Wittgenstein's point is exactly this kind of reversal of conceptual priority. He writes:

> A language-game is the ultimate arbiter of rule-following. It is therefore conceptually prior to rules, including the rules which we would naively think as defining it. This is the major conceptual revolution which Wittgenstein is attempting to carry out in the rule-following discussion. (ibid.)

Thus, for Hintikka, Wittgenstein's target with his remarks is both a false idea about the "phenomenological character of rules and rule-following" and the notion that any symbolic expression of the rule is "essentially connected with rule-following" (ibid.). Hintikka concludes:

> He [i.e. Wittgenstein] does not mean that we do not often, perhaps most of the time, make use of the symbolic expression of a rule in following it. Rather, he is arguing that this is not what following a rule *consists in*. What is [sic] consists in is being a "move" in an entire language-game, which thus is the only "criterion" of rule-following. (ibid.)

That is to say, to explain why a certain expression of a rule, e.g. '+2' means what it does is not essentially to make an appeal to the rule that this expression supposedly stands for, but rather to the whole language-game in which it is found (cf. e.g. §31, where Wittgenstein discusses a case of someone learning how to play chess).

And put in this way, it is obvious why Wittgenstein chose to frame his discussion in terms of following an explicit rule: if learning the meaning of a word is a question of learning how to follow a rule, how can that be the case when the word signifies a rule? The inevitable conclusion is either a regress: a rule is required to learn how to follow a rule, and so on *ad. inf.*, or a circle empty of content: to learn how to $\varphi$, one must learn how to $\varphi$. The easiest way out of this conundrum (as evidenced by the difficulties evident otherwise) is simply to reject that rules are

the most fundamental phenomenon in this process: to reject (RG) and adopt a different view, where rules are not the most basic phenomenon. Viewed in this way, Wittgenstein's argument even becomes surprisingly simple.

The trouble with Hintikka's reading, however, as I argued in Chapter 1, is of course that it isn't clear at all what it means to say that the language-game (or our practice, which I take to be equivalent) is what gives the rule its correctness conditions in the first place. We find it quite easy and intuitive to understand what it would be for a rule to settle whether something is correct or incorrect, but that is not the situation with language-games. Wittgenstein has told us that this has to do with our agreement in judgements, mastering of techniques and so on, but it is unclear from his discussion what this really amounts to.

In the next chapter, I will take up this challenge and provide an account of meaning and rules that takes our practices to be primary, but nevertheless avoids those problems—and offers a clear way of explaining (right or wrong) how our linguistic practices can provide correctness conditions for our use of language.[16]

16. Before moving on to the next section, I would like to make a short comment about a remark in *Remarks on Colour* where Wittgenstein seems to explicitly say that language is a rule-governed activity. The remark in question is as follows (*RoC*, §303):

> The rule-governed nature of our languages permeates our life.

Here, one might think that this remarks shows without a doubt that Wittgenstein endorses (RG)—what else can the 'rule-governed nature of our language' mean, if not an endorsement of (RG)? However, this remark is based on an translation error. In the original, the remark reads:

> Die Regelmäßigkeit unsrer Sprache durchdringt unser Leben.

where 'Regelmäßigkeit' is best translated as 'regularity', not 'rule-governed nature'. This remark can therefore not be used to argue that Wittgenstein endorsed (RG).

# Chapter 4

# Meaning and rules as constitutive practices defined by correlated equilibria

Although Wittgenstein (and Kripke) gives the paradox in the form of a problem about rule-following, it really is about the determination of meaning or content by finite means, as his presentation of it shows:[1] we learn the meaning of words by seeing a finite set of 'exemplars' which we are then expected to project into indefinitely many new cases—by seeing them as the same. For example, if we have seen a set of green exemplars—things which are supposed to exemplify the concept *green* for us—we are only competent in applying the concept *green* if we can apply it to green objects we have never seen before and not to things that are not green. However, as the paradox makes clear, any set of finite examples can accomodate any new exemplar, if the sameness relation is tinkered with in the right way—if we have a set of things we would call 'green', a blue thing might fit in the set after a specific time, if the actual concept being taught turned out to be the concept *grue* fixed such that grue things are blue after that time, and green before.[2]

It is no use simply to say that the concept we are in fact using is the concept

1. As we saw in Chapter 1, he presents the problem as a dialogue between a teacher and a pupil, where the latter sees a finite number of examples and is then expected to go on.

2. This way of putting the matter is heavily influenced by Kusch 2002, pp. 202-203 and Bloor 1997, pp. 9–14.

*green* and not *grue*, and thus that a blue thing could never fall under that concept, since whether or not the word 'green' picks out the concept *green* is precisely what is at stake. The relation between the word 'green' and the concept *green* is contingent, and we are all in the same boat, as all of us, thought of individually, or as a group, have only ever encountered a finite sample of the things that exemplify any of our concepts. Anything we've ever done in using, explaining and teaching the meaning of the word 'green' is such that infinitely many grue-like concepts fit with it in the future and so, any blue object could be said to fall under the concept we take ourselves to be using now (that we *call* 'green'), if we just give the right sameness relation between that object and the set of the things correctly described as green up to now—and crucially, we've also ever learned the concept 'the same' by seeing a finite example of exemplars and are expected to project *that* into new and new cases. It is therefore also underdetermined what our concept of 'the same' is.

It then follows, if the meaning-sceptic is right, that we cannot really say that we are making a mistake if we'd call a blue thing 'green' as there is nothing that distinguishes our referring to the concept *green* by our word 'green' instead of some *grue*-like concept, where it would be correct.[3] Anything we might do in the future fits with what we have done so far *on some conception of what counts as doing the same thing as before*, as extrapolating into new and new cases depends on what we have done previously—continuing in the same way, and there is no way to specify in advance what 'the same' is independently of the correctness of our application in that case.

Despite these considerations, however, we do not doubt that there is a fact of the matter whether or not an object nobody has ever seen before and not a part of the exemplars so far encountered could be classified as green or not—that it either is or is not green. We think that there is such a thing a *really* applying a concept in the same way as we did for our exemplars of it and the sceptic's demand is for something that picks out one unique sameness relation from the past to the future as the correct one—as *really* be doing the same thing in that case. In short, we want something—a fact—that can provide *objectivity* in our use of concepts and

3. This way of setting up the problem might seem to be putting too much emphasis on ostensive definitions, rather than verbal teaching. I don't think there's any loss of generality here. Even if we complicate our notion of how we learn concepts, it will still be finite.

an explanation of how meaning one thing now can provide this objectivity in the future before we actually employ the word.

The sceptics demand, therefore, is for something that picks out one unique sameness relation from the past to the future as the correct one—as *really* be doing the same thing in that case. If we have that, then we have a way of saying that an object nobody has ever seen before and not a part of the exemplars so far encountered could be classified as *green* or not—that it either is or is not *green*, and therefore we could explain why some utterances are correct and why some are incorrect, and hence get the correctness conditions we need.

In this chapter, I give a solution to the rule-following paradox that tries to solve the problem from this angle—in terms of what I will call *basic constitutive practices*. I will argue that by defining such practices using a game-theoretic framework, we can explain how one relation from past cases to the future can be picked out as being correct. This solutions will rely on crucial Wittgensteinian concepts, and while it is not Wittgenstein's, it is intended to answer the question posed in Chapter 1, namely how it can be that rule-following and meaning are ultimately practices, grounded in training and agreement.

## 4.1 Desiderata for our solution

In the preceding three chapters, I've mentioned and discussed some problems that a solution to the rule-following paradox must be able to solve, as well as a number of desiderata and constraints that it should meet. In order to make the evaluation of the solution presented here a little bit easier, I give an overview of this discussion here, before giving the solution itself.

First of all, we want to be able to account for the correctness conditions of meaning and rules, e.g. first of all to satisfy our principle (C):

**(C)** At each step in the application of a rule, there are certain actions that are correct and all others are incorrect, independently of any judgement or belief any particular agent might have about it.

However, since we want to reject the doctrine that language is governed by rules, we also need to satisfy some analogue of (C) that does not take rules to be pri-

mary. I will take the following principle, related to the principle (C′) discussed in Chapter 2, to be that analogue:[4]

**(MC)** If 'F' means *F*, then for any application of 'F' to an object *x*, either *x* is *F* or *x* is not *F*, independently and in advance of anyone's judgement that it is so.[5]

The terms 'correct' and 'incorrect' do not appear in this principle, but the relation it has to correctness conditions is fairly straight-forward: if an agent *S* means *F* by his use of the term 'F', then *S*'s use of the term 'F' when applied to an object *x* is correct if and only if *x* is *F*. Or perhaps more simply: saying that '*x* is *F*' is correct if and only if *x* is *F*.

The solution I will develop in this chapter is a community solution. Nevertheless, we would want this principle to be able to make room for the whole community to make a mistake, to not equate what everyone says with what is correct. In other words, there should be a tripartite distinction between what *S*'s judges to be the case, what *S*'s community judges to be the case and what really is the case.

Second, we are looking for something that constitutes the meaning of *S*'s utterances—something that explains under which circumstances meaning statements such as "*S* meant *p* by 'p'" are true. Here we would want to be able to avoid the problem of error and be able to explain linguistic mistakes. The former is the problem of 'overfitting' our principle (MC) by making *everything S* might be disposed to say be correct, relative to some rule: if *S* says '5' he meant *quaddition* and if he says '125', he meant *addition*, and so he is always right. The second is the problem of *S intending* to say that *x* is *F* and being correct in fulfilling this intention, even if *x* is not *F*.

Third, we want an account that does not place too heavy epistemological burdens on the agent, and account for the feeling of guidance by the meaning or the rule. That is to say, we want to be able to explain how an agent is able to understand a rule or the meaning of a term and subsequently do what the meaning or the rule require.

4. That was, recall, the principle that if *S* means *F* by his use of a symbol $\varphi$, then it is correct for *S* to apply $\varphi$ to an object *x* iff *x* is *F*.

5. (MC) from (C) and ' Meaning'.

Fourth, we want to be able to make a distinction between merely acting in accord with a rule and properly following it, as well as be able to say what it is to be justified in making an utterance or apply a rule to a new case.

**Wittgensteinian considerations** In Chapter 1, I argued that Wittgenstein's own remarks are not sufficient to give us a satisfactory solution to the paradox, and that even if we appeal to our common form of life, mastering of techniques and our practice of following rules, it is still unclear how those obscure concepts can help. In presenting the solution here, I try to make these concepts do the work required, and as such, the account proposed here is definitely Wittgenstein*ian* in that narrow sense. I will not, however, argue that the solution presented here is Wittgenstein's. It decidedly is not. Rather, it is inspired by Wittgenstein's concerns and Wittgenstein's discussion of rule-following, and is intended to serve our goal: to defend Wittgenstein's radical conventionalism about mathematical truth in a way that contemporary analytic philosophers might understand and appreciate.

There is one last desideratum that I would like to flag, before getting on with it. It is common among interpreters of Wittgenstein to argue that his rule-following paradox shows that nothing outside our practice of using a term can secure the meaning of that term—that a rule and the practice of following it cannot be separated.[6] I will claim that the solution presented in this chapter can give us a good way to understand such, otherwise obscure, claims.

## 4.2 Meaning as a constitutive practices

In this section, I will present my own account of rule-following and meaning, using the game-theoretic resources of a recent account of convention by Peter Vanderschraaf as my framework.[7] I will suggest that constitutive rules and meaning are *basic constitutive practices* and that the framework of Vanderschraaf's theory of convention can provide one way of seeing how it might be possible for such practices to have enough structure to solve the problem without regress.

6. Authors that we've looked at so far which have made this claim are e.g. Verheggen 2003; Hacker and Baker 2009.

7. See Vanderschraaf 2018.

I will first explain the notion of constitutive practice that I'm working with, then give the details of Vanderschraaf's account that I require and then give my own account. I will *not* argue that constitutive practices *are* conventions as such, but merely help myself to this particular game-theoretic framework in explaining their structure. They will turn out to have certain features of conventions, but I will stop short of claiming that they are here. I will discuss this further is Chapter 7.

In order to reduce the burden on the reader as much as possible, however, I will try to explain my account with a minimum of formal details, and instead refer to the appendix of the dissertation for the full formal story, as well as Vanderschraaf's book.

**Basic constitutive practices** In Chapter 3, we discussed the distinction between *regulative* and *constitutive rules*. The former kind, we said, are rules that regulate a practice whose existence is logically prior to the rules that regulate it and the latter *constitute* an activity whose existence depends on and cannot be explained or described without reference to those very same rules. Constitutive rules, as Searle puts it, can therefore be said to "create and define new forms of behaviour" (Searle 1969, p. 33). The important idea for us here is that constitutive rules *constitute* some practice—make that very practice possible. The rules of football, for instance, make it possible to perform certain actions that did not exist prior to the rules being in force and having attitudes towards those actions, e.g. scoring a goal and intending to score a goal. Furthermore, what it is to play football is explained by reference to those rules. On this view, football is a constitutive practice, because it is constituted by its rules.

While this distinction seems intuitive enough, there are certain problems in cashing it out in terms of descriptions. One of these problems, noted by Searle himself, is that any regulative rule can be re-described as a constitutive one, since any regulative rule $R$ describes a new form of behaviour, namely that of following $R$—thereby erasing the distinction. If constitutive practices are those that are constituted by a rule, then any regulative rule is also a constitutive one, just constituting the practice of following itself.[8] If it turns out, however, as I will argue, that

8. See e.g. Reiland 2019; Glüer and Wikforss 2009; Marmor 2009; Ruben 1997 for further discussion on this and other problems with Searle's formulation of the distinction.

rules are constitutive practices themselves, this result of the distinction will not be a problem to be accounted for, but rather an expected feature.

Since I want to explain rules as constitutive practices, however, I cannot rely on a specification of what that is which essentially relies on rules. I will take my cue from Rawls instead, who suggests that actions (or rather action types) that belong to a constitutive practice cannot be performed outside of what he calls the 'stagesetting' of the practice (Rawls 1955, p. 27) and hence a constitutive practice is one that makes such actions possible. For instance, it is impossible to score a goal outside the stagesetting provided by some game where goals are scored. Football is, on this view, a constitutive practice, because the only way to evaluate some action as 'scoring a goal' is by reference to the practice of playing football—and in this case, the rules are what provides the necessary stagesetting. Furthermore, an agent engaged in such a practice can only justify and explain his own action by reference to the practice that constitutes it, since the action has no description outside the practice.[9]

Accordingly, I will say that *P* is a *constitutive practice* if it is only possible to say what it is to take part in *P* by reference to some kind of stagesetting that defines what *P* is—something that constitutes *P*. This definition is admittedly quite vague, but should fit the intuitive examples. For example, on this definition, chess is a constitutive practice because it is impossible to play chess without the rules that define the game and impossible to describe an action within the game without reference to the practice of playing it. In this case, it is the rules that provide the necessary stagesetting. The practice of driving on the left side of the road, on the other hand, is not a constitutive practice, because it is possible to drive on the left side of the road without any stagesetting. Notice, however, that I do not require that every practice which is constituted by constitutive practices to be itself constitutive. The practice of playing chess only with wooden pieces is not a constitutive practice nor is playing football every Wednesday, for instance, even if those practices are constituted by constitutive practices.

Further, I will say that *P* is a *basic* constitutive practice if it is a constitutive practice and does not require some further constitutive practice as stagesetting.

9. This is a point made by Rawls. Cf. also *Philosophical Investigations*, §381: "How do I know that this colour is red?—It would be an answer to say "I have learnt English"."

The idea that I will develop in the rest of the chapter is that constitutive rules (and meaning) are basic constitutive practices: to follow a constitutive rule *R* is to take part in the basic constitutive practice of following *R* and to mean *addition* by the symbol '+' is to take part in the basic constitutive practice of using the symbol '+'. Every rule, therefore, has an associated basic constitutive practice which determines what it is to follow that rule. Chess, on the other hand, is not a basic constitutive practice, since it requires rules for its stagesetting, which in turn are basic constitutive practices.

To prevent misunderstanding, I should stress that I am not claiming that constitutive practices are governed by constitutive rules, nor that meaning is constituted by such rules. Rather, the claim that I will defend for the rest of the chapter is that meaning and constitutive rules are *both* explained by what I call constitutive practices. In fact, there is no analog of regulative rules on my account as a counterpart to constitutive practices, i.e. no such thing as regulative practices. In the terms used in the last chapter, it will be the case that for any concept *C*, there is a rule *R* which has the same correctness conditions as *C*, and maybe used by the agents in explaining their own practice, but this rule *R* does not govern *C* and is not the source of *C*'s correctness conditions.

What is left then, is to account for basic constitutive practices, those that require some kind of stagesetting that define what they are, but do not look outside themselves for this feature, so to speak. I will argue that the game-theoretic *structure* of such practices is what can provide this stage-setting.[10]

**The meeting game: an example** We will, following Vanderschraaf, assume that agents engage in a sequence of interactions—coordination games—where each agent can perform some action based on their beliefs about the actual world and receive some kind of pay-off depending on what the actual world in fact is. The pay-off they expect to receive is based on the actual pay-offs in each case and

10. I'm by no means the first to try to apply game-theoretic methods to the paradox. Giacomo Sillari has for instance tried to use Lewis's theory of convention to the same effect (Sillari 2013).

Francesco Guala and Frank Hindriks have likewise presented an influential account of social ontology that presents institutions as correlated equilibria to which constitutive rules can be reduced (Guala and Hindriks 2014). As far as I can see, however, they take regulative rules for granted in their account, which begs the question if the problem to be solved is the rule-following paradox. This is of course by no means criticism of Guala and Hindriks, since that is not their goal.

their beliefs about how likely it is that they will get it. We will leave it quite open what the pay-offs actually are, but they do not have to be anything tangible that the agents get, nor do the agents have to be selfish in order to prefer one thing to another. They might prefer to help others, for instance, and thus receive a higher pay-off in situations where others are helped. Pay-offs might even be thought of as external to the agents, as just marking an occasion of successful communication, or the like. We say that each interaction is in equilibrium if no agent could unilaterally change their move and receive the same or higher pay-off. In other words, given what everybody has done, nobody should nor would do anything else.

To illustrate, let's start with a simple example of a coordination game, taken from Lewis (D. Lewis 1969). Suppose two people, let's call them $R$ and $C$, have a desire to meet.[11] They do not really care where they meet, so long as they do. $R$ and $C$ must both decide where to go. The best place for $R$ to go to is where $C$ will go and vice versa. Each person will thus choose relative to where they think that the other person will go, and if one of them succeeds, they both do.

Let each of the places $R$ and $C$ can go to be denoted by some natural number $n$, and a tuple $(Rn, Cm)$ stand for the possible combination of choices $R$ and $C$ can make. For example, $(R1, C2)$ would mean that $R$ went to the place denoted by 1 and $C$ to the place denoted by 2, and so on. Each agent gets assigned a pay-off based on the actions both of them took (maybe representing the enjoyment they get from having dinner together, if they succeed) and since in our case, both agents are happy if and only of they meet, we can stipulate that their pay-off in this case is 1, and 0 otherwise. If there are three possible places to choose from, we can then represent all possible combinations of actions and their pay-offs by a matrix, where the first number in the ordered pairs represents the pay-off of $R$ and the second that of $C$ (see Fig. 4.1).

We say that an *equilibrium* is a combination of actions where neither agent can do any better, given the action of the other agent. No player then, would opt to change their action, if they knew what the other agent would do. If the game is not in an equilibrium, an agent would make such a change. For instance, if $R$ chose 1 and then finds out that $C$ chose 3, $R$ would now prefer to choose 3. In a certain sense then, the equilibria are the only combinations of choices that are stable—if

11. "R" for "Row" and "C" for "Column".

| | $C1$ | $C2$ | $C3$ |
|---|---|---|---|
| $R1$ | $(1,1)$ | $(0,0)$ | $(0,0)$ |
| $R2$ | $(0,0)$ | $(1,1)$ | $(0,0)$ |
| $R3$ | $(0,0)$ | $(0,0)$ | $(1,1)$ |

Figure 4.1: Meeting game

any agent would unilaterally change his move, he would be worse off.

If we suppose that the agents keep meeting each other regularly, we can define a sequence of games $(\Gamma_t) = \Gamma_1, \Gamma_2, \ldots$ where each $\Gamma_i$ has the same form as the meeting game above. We call such a sequence a *supergame* and the index $t$ a *period*. Since the agents know the history of their interactions and we assume that they have beliefs about what the actual world is like, the following strategy is available to them:

$f_i$: If we haven't met, choose an arbitrary place; otherwise go to the last place we met.

Since no player could do any better than $f_i$ until they meet for the first time, and certainly not afterwards, the strategy system $\boldsymbol{f} = (f_R, f_C)$ is an equilibrium of the supergame $\Gamma$. Notice, however, that even though $\boldsymbol{f}$ is an equilibrium, it doesn't tell us what the agents will actually do—i.e. which meeting place they will actually choose. It might be, for instance, that both agents choose to go to 1 in the first period, in which case they will, according to $f_i$ always go to 1 afterwards. It might also be that $R$ goes to 1 in the first iteration and 2 in the second, while $C$ goes first to 3 and then to 2, in which case the agents will both go to 2 in every iteration afterwards. It is therefore not until the agents have met for the first time that $f_i$ settles which action the agents will take in all cases after that.

We say that the combination of the actions each agent takes at a given period $t$ is an *act profile* $\boldsymbol{s}_t$ of that period. If the actions are taken in accordance with a strategy which is in equilibrium (such as $f_i$), we call a sequence of such profiles an *equilibrium path* of $(\Gamma_t)$. In our case, the set of act profiles where the agents always meet at 1 is such a sequence, the set where they always meet at 2 is such

a sequence, and the set where $R$ goes to 1 in the first iteration and $C$ goes to 2, and both always go to 3 afterwards is such a sequence. In fact, the equilibrium $\boldsymbol{f}$ defines infinitely many equilibrium paths through $(\Gamma)$, since before our agents meet for the first time, there's always a chance that they won't in the next iteration. After they do meet, however, they immediately settle what the path is from that point onwards.

Notice, however, that the real content of the strategy $f_i$—the actual actions taken in accordance with it—is not given until an equilibrium path is fixed, and it is therefore not necessary that the agents have any particular path in mind when they choose a strategy, and therefore no particular actions in individual cases either. This property of equilibrium paths will turn out to be vitally important for the solution on offer.

### 4.2.1 The present account: '+' games

My intention is eventually to define the correctness conditions for the use of the symbol '+' by building up the basic constitutive practice of using it in terms of games like the meeting game and the main claim is that in each case, the practice of using the symbol in the language has this structure in basic cases. This structure is the stage-setting we can evaluate the agent's actions against—and if they lie on the equilibrium, then they count as an instance of the action that the basic constitutive practice is a practice of.

I will suppose that the agents in the games I'm going to define are such that they respond to training and instruction in similar ways. In other words, that when they see or interact with a finite set of exemplars for each concept they learn in their linguistic training, they are so endowed that the react to new cases in a uniform manner. After such training, the agents form dispositions to use the words they are taught, i.e. to judge that a certain object falls under a concept or to continue a calculation in a particular way.

Accordingly, I will also stipulate that whenever an agent has the disposition to use a term in a certain way, then the agent also believes that this way is the correct way (or would, if they had any belief about that particular case). In particular, *i*'s dispositions about sum-language and his beliefs about sums never come apart. As

an example, if the agents are anything like me, they will have formed the disposition to reply '125' and not '5' to the question what is the sum of 57 and 68 and likewise have formed the belief that giving this reply is what 'adding' these two numbers is (but see below for discussion of the content of *S*'s belief in this regard).

I will start by considering what I will call the 'simplified '+' game'. This game is played by two agents and their pay-offs and possible moves are given by the matrix below. I will imagine that the simplified '+' game is played indefinitely many times

| | '5' | '125' |
|---|---|---|
| '5' | 1, 1 | 0, 0 |
| '125' | 0, 0 | 1, 1 |

Figure 4.2: Simplified '+' game for two agents

by the agents and at every period each player can give one of two responses to a question of the form 'What is 57 + 68?'. Each period stands for every possible use of the symbol '+'. This is supposed to reflect the fact that there are indefinitely many uses of the symbol '+' when it is flanked by the numbers 57 and 68 and we want to make sure we have enough of them to cover any case. The simplified '+' game for two agents is therefore really a supergame of the form $(\Gamma) = \Gamma_1, \Gamma_2 \ldots$.

Given our assumption that (a) agents follow the strategy of replying with what they judge to be the correct answer, and (b) that their dispositions to judgement are very similar, the simplified '+' game will only have one equilibrium, namely the one where the agents give the answer they believe is correct, and only one equilibrium path (and if the agents are anything like us, that path will be defined by the answer '125'). Since the beliefs of the agents are not probabilistically independent (they form them after similar training), we call this a *correlated equilibrium*.[12]

**Stable dispositions to judgement** The 'reactions' or 'dispositions to judgement' that act as inputs for the game should be something more than just 'brute'

12. This solution concept is more general than the Nash equilibrium and was first studied by Robert Aumann (Aumann 1974, 1987). See also Vanderschraaf 1995, 1998, 2018 and Gintis 2009 for discussion.

responses, however, because the way we actually use words is quite a bit more complex than the simple model considered here might imply. This is also reflected in our linguistic training. For instance, I might be disposed to judge that two lines are of unequal length if they are shown to me by means of a clever illusion, but not so disposed if I place a ruler on top of them and read the same number off the ruler in both cases. Likewise, if I regularly write that '68 + 57 = 5' (by some systematic slip of the pen) but accept correction by my calculator or a helpful and patient friend, we might say that I'm disposed to say that $68 + 57 = 5$ but also that $68 + 57 = 125$. Here, I have two sets of dispositions towards the same pair of lines and two sets toward the same calculation.

If we want to account for meaning as basic constitutive practices in this way, we need to take such considerations into account. The most obvious way to do this is to make only certain dispositions 'count'—by stipulating, for instance, that only dispositions under certain conditions $C$ are meaning-determining, perhaps where $C$ denotes 'standard' or 'ideal' conditions. The trouble with this approach is twofold: there is (a) a risk in making the account circular, by defining our dispositions to add as those dispositions we have under $C$, where $C$ is then defined as those conditions where we are disposed to add; and (b) a problem of explaining why $C$ are the right conditions, and not some other conditions $C^*$. If we only adopt $C$ because they give us the right dispositions, our solution is both *ad hoc* and perhaps also circular.[13]

I believe we can avoid both these problems, however, by relying on our assumption that the agents' meaning-determining dispositions are formed by their linguistic training. This training is such that $S$ not only forms first-order dispositions to judgement about individual cases, but also higher-order dispositions to judge about his own judgement—'dispositions about dispositions to judgement'. Suppose for instance that $S$ has learned to use colour words by being shown examples, being corrected when he makes mistakes, etc. If $S$ were to inspect a necktie under strange electric lighting, for example, $S$ might then judge that the tie is blue, but outside on a clear day, $S$ might judge that it is green. Here, because how $S$ has learned to use the words 'green' and 'blue', $S$ would be disposed to judge that his first judgement was incorrect, but disposed to stand firm with regards to his sec-

13. Thanks to Andrea Guardo for this point.

ond judgement. This follows from the very way we assume *S* learnt to use colour words.[14]

Accordingly, I will stipulate that only *S*'s dispositions to judgement count as determining an equilibrium of a constitutive game of a supergame ($\Gamma$) which are *stable* and the agent is, in some sense, not prepared to withdraw. This is not circular, because I do not define *S*'s dispositions in terms of standard or ideal conditions, for example, but say that standard or ideal conditions are those in which *S*'s dispositions are stable. The explanation for why *S*'s dispositions are stable in the latter case and not the former is simply that *S* has acquired the relevant higher-order dispositions through his linguistic training.[15]

In this sense, the way we come to learn concepts is constitutive of the meaning of the words we use to refer to them (which is a natural consequence of the account, given what we've already assumed). Similar considerations would lead us to conclude that only my dispositions to judge that $68 + 57 = 125$ are stable, as well as those concerning the length of the two lines when I place a ruler on top of them, but not the ones I have when I just rely on my visual impression.

**The generalised '+' game** Now, the simplified '+' game is only defined for two agents and two possible responses for each agent and only concerns a limited use of the symbol '+' (but nevertheless for indefinitely many occasions of use). We can generalise it by allowing indefinitely many agents in line with our definitions above and allow the agents infinitely many replies. By letting the numbers go variable in 'What is 57 + 68?', we can then define a generalised '+' game for any $n$ and $m$ in the question schema 'What is $n + m$?' such that the game allows indefinitely many agents and indefinitely many replies in each case.

Since there is a countable number of generalised '+' games, we can enumerate them e.g. as follows

$$(\Gamma)_+ = (\Gamma_t)_1, (\Gamma_t)_2, \ldots$$

where, recall, each supergame in the sequence is indexed by period $t$, and is thus

14. This example is from Sellars (Sellars 1997, pp. 37–39).

15. This is reflected in Wittgenstein's original exposition of the paradox (*Philosophical Investigations*, §§186–188): it is vital that the student does not accept any corrections, but keeps insisting that what they did is correct and can rationally do so.

repeated indefinitely many times. Each of the supergames in the sequence has an associated a set of possible equilibria, each given by the possible dispositions of the agents and defined by a unique equilibrium path. If we were to choose one equilibrium from each supergame in the sequence, we'd have a set of equilibrium paths corresponding to one possible sceptical interpretation of '+'. For instance, to get the set of equilibrium paths corresponding to *quaddition*, we could choose the equilibrium corresponding to *addition* for every generalised '+' game up to $n = 57$ and the equilibrium path given by $\boldsymbol{s}_t = \{(\text{'5'}, \text{'5'}, \ldots)\}$ in any game afterwards. Call such a selection a *second-order equilibrium path* through $(\Gamma)_+$. We can think of such an equilibrium path as one possible interpretation of the term '+' or, alternatively, say that the extension of any possible concept in a given case is given by some such second-order equilibrium path.

Given our assumption that the agents form a similar set of dispositions regarding the use of the symbol '+', there will only be one equilibrium path through each of the supergames in the sequence $(\Gamma)_+$ and hence only one second-order equilibrium path through $(\Gamma)_+$ itself. There will be one equilibrium path through $(\Gamma_t)_1$, one through $(\Gamma_t)_2$ and so on, for each supergame in $(\Gamma)_+$. This is, again, because of our assumption that the agents form similar dispositions to judgement and follow the strategy of replying in accordance with their beliefs about a given case. The agents are not guessing or deducing what the equilibrium will be, however, as from their perspective there is always only one possible option: the one that fits with their stable judgement about a given case, as that is what they will believe adding is. As such, the practice itself is opaque to the agents taking part in it.[16]

The second-order equilibrium of $(\Gamma)_+$ therefore represents the structure of the agents' actual practice regarding the use of the symbol '+'. The structure of the practice itself can then act as the stagesetting required for us to evaluate action as being an instance of that practice without circularity and hence we can say that what it is for the agents to be taking part in the practice is to perform the action that lies on that second-order equilibrium path—to be correctly 'adding' in the case of $(\Gamma)_+$.

16. The obvious concern here, given the discussion in Chapter 2, is that we do not have the necessary dispositions—that my account relies on dispositions that cover infinitely many cases, and those are simply not available. I discuss this problem below.

The second-order equilibrium path defines what it is to take part in the basic constitutive practice of using the symbol '+' and therefore also what counts as doing the same thing as in a previous case. In other words, what it is to be adding is to be taking part in the basic constitutive practice of adding and the correct answer in each case is given by the structure of the practice and the dispositions of all the agents. This is the fact that the sceptic is looking for—i.e. what picks out one sameness relation from a set of exemplars to future cases as *correct*, and it does so by selecting it as being constitutive of the practice itself. The agents do not need to have any particular action in mind for the practice to settle on a given answer as correct, and can rely only on their dispositions and beliefs about what is correct in a given case in choosing what to do.

Accordingly, *S* meant *addition* by his use of the symbol '+' in the past and not *quaddition* because *S* was taking part in a basic constitutive practice of using the symbol '+' defined by the actual second-order equilibrium path of $(\Gamma)_+$ and the particular equilibrium of the supergame that deals with 57 + 68 does not allow '5' as a correct answer in this case, only '125'. If Kripke's sceptic were to turn around—as so often—and ask what fact makes it the case that *S* is just not taking part in the constitutive practice that sanctions the response '5', namely the basic constitutive practice that corresponds to the equilibrium path of *quaddition*, we have a ready answer: there is simply no such practice for *S* to be a part of. There is a possible such practice, but it is not actual, given the dispositions of the other agents.

Since all the possible second-order equilibrium paths of a basic constitutive practice of using a term, e.g. 'red', have the infinite structure that they have, each of them represents the extension of a possible concept. It is the job of the practice to pick out one of these paths as correct, and hence the meaning of the term 'red' for the agents taking part in the practice. If, therefore, an agent taking part in a basic constitutive practice of using the term 'red' and the philosopher reasoning about the meaning of that term are one and the same (and we further assume that this practice is *our* practice), then the concept that the practice picks out is *red*.

**The content of *S*'s intentions** There should, however, be a distinction between *S* meaning *addition* by his use of the symbol '+' (or the word 'plus') and *S*'s other use of the symbol—*S*'s aimless doodling, perhaps, or mindless parroting of

arithmetical propositions. Here, we should first distinguish between the meaning of a sentence in *S*'s language and *S*'s meaning at some particular occasion. The meaning of '+' in *S*'s language is here explained by the structure of the practice *S* is embedded in and we do not require anything further. *S*'s mechanically repeating arithmetical statements does not detract from their meaning in this case.

In the other case, I do not see any viable candidate to make this distinction other than *S*'s *intention* to mean *addition* by his use of the symbol '+'. It is commonplace in discussions of the paradox to take it to show that intention is somehow not an occurrent mental state. The suggestion that I would make here is that the content of *S*'s intentions are not fully specified by *S*'s mental state, but rather by the content of *S*'s dispositions to use '+' *and* the structure of the basic constitutive practice of using that symbol, of which *S* is a part. *S* can therefore intend to utter a sentence with a particular meaning on a particular occasion, but the intention doesn't independently run ahead on the 'rails to infinity' and settle every case without reference to the practice *S* is embedded in. *S* intends to *add* and has certain beliefs about what counts as adding in a given case—but the full content of the intention cannot be specified without reference to $(\Gamma)_+$.

We might therefore (crudely) say that *S*'s mental state somehow tokens the term '+' and its actual content depends on the practice of using '+'—the practice of adding. For instance, we could imagine (again, very crudely) that *S*'s mental state of intending to say that *p* consists of him saying to himself: 'I'm going to say that *p*', in which case the very fact that *S* is embedded in the basic constitutive practices that gives the terms in the expression 'p' its meaning *enables S* to say that *p* and gives his utterance that tokened 'p' its content—the intention to *p*.[17]

Since the same point would apply, *ceteris paribus*, to all other contentful mental states that *S* might have, for instance his beliefs, his wishes, his understanding, etc., this view is a kind of *social externalism* about content.[18] Two agents, *S* and *T*, might have the exact same mental states, for example, to use our crude model again, both saying to themselves 'I'm going to say that *p*', but these states will not have the same content unless *S* and *T* are both taking part in the same basic constitutive

17. I do not mean that this crude picture is correct. It is just intended to give an idea of what kind of account of content follows from the view being developed, however we want to actually spell out such a view in the end. The same point would apply to a more sophisticated view.

18. The *loci classici* are Burge 1979, 1986.

practice.

This already explains how *S* can misapply a word: *S*'s intention, in a case where he makes a calculation mistake, includes the token 'addition', but his actions do not conform with the correctness conditions of the practice of adding from whence the intention gets its full content. That is to say, the content of *S*'s intention to add is specified by $(\Gamma)_+$ and not merely *S*'s dispositions or actual utterances. If *S* were then to make utterances that miss the equilibrium, that would not show that *S* had a different concept, hence avoiding the problem of error.[19]

**A genuine case of rule-following** On this account, there is a sense in which there is a constitutive rule in force for *S*—the one we can trivially read off the set of equilibrium paths of the practice and gets its correctness conditions from the practice. In this case, that might for instance be the constitutive rule

'$x + y = z$' is correct if and only if $z$ is the result of adding $y$ to $x$.

The correctness conditions of this rule are settled by the practice, because the meaning of they symbol '+' (for the agents taking part in the practice) is given by $(\Gamma)_+$ and likewise the meaning of the word 'adding' for those agents.

In Chapter 2, I criticised Warren for treating the problem as if we had a metalanguage at our disposal in which all the relevant concepts can be taken for granted. I argued that the problem is about how meaning is constituted in the first place, and that includes for us, the theorists reasoning about *S*'s use of language. We can now see how this account avoids that circularity/regress problem. If we assume that we, the philosophers reasoning about *S* meaning, are also participants in the same basic constitutive practice of using the term '+' as *S* is, the meaning of this rule for us, as well as *S*, is determined by that practice. In that sense, the practice determines the meaning of "adding" as well as the meaning of '+', and as such, the basic constitutive practice is active on both sides of the biconditional. We

19. This is one way to understand Wittgenstein's answer to his own puzzle about intention: What makes it the case that *S*'s intention is of that which is intended? How can getting an apple be the fulfilment of wanting an apple? Wittgenstein's answer: "It is in language that an expectation and its fulfilment make contact" (PI, §§437–445).

See also (PI, §337): "An intention is embedded in its situtation, in human customs and institutions. If the technique of playing chess did not exist, I could not intend to play a game of chess".

are, as McDowell puts it, as much involved on the left-side of a truth-conditional bi-conditional as we are on the right side (McDowell 1984, p. 352).

The rule, however, is not really motivating *S*'s actions in any substantial way and therefore it would be misleading to say that *S*'s use of the symbol '+' is rule-governed. For any basic constitutive practice, there will be such a rule, but it is not governing that very concept. But since *S*'s mental state is a part of what explains what *S* meant and the structure and role of *S*'s practice in giving those mental states content is not apparent to *S*, it would *seem* from *S*'s perspective that *S* is being guided by his meaning, rather than mere dispositions derived from his training. The feeling of being guided by meaning is therefore an illusion, explained by the overall structure of the account.

In what I've said so far, it might then seem that there is really no such thing as rule-following proper. After all, in the cases I've considered, *S* is not motivated by the rule that constitute his practice at all—the rule is a constitutive one, defined by the structure of the practice *S* is engaged in and does not play any motivating or causal role for *S*—in the sense that the rule is what makes *S* do one thing rather than another. But it is a truism that if *S* is following a rule, then the rule *S* is following must play some role for *S* in what he does, and further that there is a distinction between merely acting in accordance with a rule and to be really following it. What then is rule-following?

In Chapter 3, I discussed one plausible model of rule-following, Crispin Wright's 'modus ponens model' of rule-following. Consider again Wright's example of a basic case of a rule-governed use of language where *S* is using some basic predicate, such as 'red'. In this case, the rule *S* employs and how he does it, would be something like the following:

(Rule) If …*x* …, then it is correct to predicate 'red' of *x*.

(Premise) …*x* …

Conclusion) It is correct to apply 'red' to *x*.

As Wright points out, if we want to include cases like this in the modus ponens model, we require an 'anterior concept' which determines whether or not the right conditions obtain for the application of the rule, namely the one indicated by '…*x*

…'. If 'red' really is basic, this concept cannot be anything else than *red* and so the ability of the rule to guide *S*'s actions seems to have evaporated.

The basic dilemma was, that we either need to abandon the idea that language is at bottom rule-governed or find a different model of rule-following, a less natural one, perhaps, than the modus ponens model. On the account offered here, we can opt for the first horn: *S*'s use and understanding of a basic predicate like 'red' is not to be understood in terms of rules at all, but a constitutive practice with a particular structure. *S* is not guided by this structure in his actions, but it does give a clear way of explaining the correctness conditions (and therefore content) of the concept *S* is employing: it is given by the correlated equilibrium of the sequence of games representing that structure. There is, however, a constitutive rule in play, but one that is trivially read off the equilibrium path of the sequence of games and does not figure in *S*'s use of the concept at all. The modus ponens model is therefore inappropriate for *S*'s use of the concept *red* for the simple reason that such a basic concept is not rule-governed at all.[20]

Now consider another case of rule-following—this time a case of practical reasoning, involving the rule "If the light is red, stop!". If we put that into the modus ponens model, we get:

(Rule) If the light is red, stop!

(Premise) The light is red.

(Conclusion) Stop!

For *S* to understand the rule and that the premise obtains, *S* needs to at least understand the concepts making up the rule and the premise, which in this case do not require any further rules to be grasped. *S* can perfectly well reason about what he should do on the basis of this rule and then act in accordance with it because of his grasp of the rule—the reasoning about the rule bottoms out in concepts which themselves are not rule-governed. In the case of the premise, for instance, *S* would judge that the light is red on the basis of his own disposition to call the light red and he is correct because of the practice he belongs to that determines the meaning

20. Glüer og Wikforss (Glüer and Wikforss 2010) draw similar conclusions from the fact that we seem to be forced into blindness by the assumption that language is rule-governed.

of the word 'red'. This, of course, only works if the agent who expresses the rule (i.e. me) and *S* belong to the same basic constitutive practice of using the word 'red'—otherwise the expression of the rule and rule actually denoted might come apart.

If we understand basic cases of rule-following to be those cases where the rule does not require *another* rule to be made sense of, this certainly seems to be a candidate. On the present account, there is therefore no particular worry about the distinction between following a rule and merely acting in accord with a rule in basic cases, since the basic cases that seemed the most problematic are now not seen as cases of rule-following proper at all. It the follows that we can account for the correctness conditions of our principle (C) quite easily, as being given by the meaning of the terms that figure in the expression of the rule, whose meaning is in turn given by the basic constitutive practices of using those terms.

## 4.3 Some objections to the account

In this section I will discuss a few possible objections to the account. I will first address some worries related to dispositions, then objections that have to do with criticisms of communitarian accounts and finally the possibility that my account is internally inconsistent.

### 4.3.1 Worries about dispositions

The shape of the present account is that agents acquire dispositions to judgement through their linguistic training, and subsequently that the meaning of a term as used by those agents is given by the game-theoretic structure of their basic constitutive practice of using the term—i.e. the second order equilibrium path through all possible uses of the term. This practice is constitutive, because the second-order equilibrium path defines what it is to be taking part in that very practice—to be adding *is* to perform the actions that lie on the second-order equilibrium path of the basic constitutive practice of using the term '+'. And since the structure and the dispositions of the agents are all given at a particular time, the second-order equilibrium path settles the correctness conditions of every subsequent use of the

symbol '+' in advance, as the inputs already cover every possible case.

It might then be objected that my account relies on dispositions that might not be available—that we simply do not and cannot have the dispositions that this account requires—dispositions that cover every possible case. In Chapter 2, we saw two arguments from Kripke to that that effect. The first argument was that it is simply not the case that we have a disposition to give the sum of any two numbers when queried: *S*'s actual dispositions are finite and to mean addition by the symbol '+', *S* would have dispositions that covered every case, but in fact, there are cases, e.g. of very large numbers, where *S* simply has no disposition or they are unreliable.

The second argument was that even if we try to repair this naïve account by complicating our notion of disposition such that *S* meant addition by '+' if *S*, *ceteris paribus*, would have responded with the sum of two numbers $n+k$ when queried in hypothetical situations where *S* has the necessary abilities and time to give a reply, then that would only work if we would already assume that *S* would reply with the sum of $n+k$ and not their quum, a question begging assumption. Kripke's challenge is, how can we spell out this *ceteris paribus*-clause without simply assuming *S* is adding and not quadding?

In our case, I think we can explain how agents can have the right kind of dispositions without making too heavy psychological demands on the agent. The first thing to notice, however, is that the focus on arithmetical examples is misleading in this respect. It is true that nobody has a general disposition to reply with the sum of any two numbers—and perhaps only dispositions to reply with the sums of very small finite numbers. Instead, our actual practice relies on calculations and certain techniques to give a reply to arithmetical questions when the numbers involved are high enough. Our practice is, we might say, mediated through a technique the mastery of which cannot be separated from the acquiring of the concept.

In Chapter 2, I criticised Warren's dispositionalist account of meaning on the grounds that it does not provide correctness conditions for meaning. Nevertheless, we can follow Warren in saying that when *S* is adding sufficiently large numbers, *S* is disposed to 'sum single digit numbers, carry and move on to the next step in the process'. There are, we could say, gaps in *S*'s dispositional table, but there is a structure to the technique and so even if *S* does not have infinite simple dispo-

sitions, *S* is still disposed to execute the first step of the algorithm, and then for each particular step *n* of the algorithm to execute that step and move on to step *n* + 1. A snapshot of how one such particular technique looks is something like the following calculation:

$$\begin{array}{rr} & {\scriptstyle 1\phantom{0}} \\ & 6\ 8 \\ + & 5\ 7 \\ \hline & 1\ 2\ 5 \end{array}$$

Performing it implicitly requires *S* to be disposed to perform step *n* in the calculation, whether it be adding small numbers or to perform the carry operation, and then be disposed to move on to step *n* + 1.

This of course just moves the problem to the concept *carry*. The sceptic can always ask what makes it the case that *S* is in fact *carrying* and not *quarrying* when *S* is performing such a calculation. After all, *S* has only ever performed a finite number of carrying operations, and what makes it the case that the next one should be performed in one way rather than another? Maybe *S* is in fact disposed to write down 2 above the numbers to the left after *n* operations and not 1. In which case, what calculation are we even speaking of?

That question, I believe, is much more psychologically tractable. We can say that *S* has a general disposition to give *some* reply when expected to carry, formed by his training, and that these dispositions in turn define a basic constitutive practice of carrying (along with the dispositions of the other agents in the carrying community, so to speak). It is not that *S* has dispositions to carry as such—conceived of independently of the practice—but that those dispositions that the agents have in the practice *S* is taking part when performing such a calculation define what 'carrying' means among those that take part in this practice—what concept the term 'carrying' picks out for the participants. It is only if *S*'s dispositions agree with the equilibrium of the practice that *we* have in using the term that we can say that *he* means *carrying*.

In our case, we can account for these correctness conditions in a non-circular way: they are given by the structure of the basic constitutive practice. The only assumption that we do need is that the agents form similar general dispositions

after being taught how to carry. We do not assume that they are in fact carrying when specifying their dispositions, we assume they have a general disposition to give *some* reply when they are in the right circumstances and say that whatever *that* is, *that* determines the meaning of the term the agents are so disposed to use, in conjunction with the structure of the basic constitutive practice. This determines what the word 'carrying' picks out in that community—and so in our case, *carrying*.

This of course requires that *S* has a disposition to carry whenever presented with a suitable case. In this case, I don't think it is necessary to stipulate an infinity of dispositions to carry nor take into account that *S* might get tired or not live long enough before his disposition to carry is manifested, and so on. The claim is rather that for any two numbers smaller than 9 whose sum is larger than 10, *S* is disposed to write 0 as the outcome and 1 above the two numbers to the left, and move on to the next step in the calculation. This claim does not place unreasonable psychological demands on the agent. That is to say, it might be implausible to think that has as a disposition to give a reply to any addition problem, but it is not implausible to think that *S* has a disposition for any carrying problem—even if that problem appears in the context of a very large number, as the carrying problem itself is always 'local', so to speak.

Consider for instance two very large numbers, for example, numbers that are so long that it would require the whole lifespan of the universe to write them down. It's clear that a finite and flawed agent would not have the disposition to go through with but a small initial segment of the calculation to add these two numbers. But it does not seem unreasonable that, for any step $n$ in the calculation, the agent has the disposition to perform the carrying operation and move on to step $n + 1$—where the content of the term 'carrying' is for *S* defined by the basic constitutive practice of carrying. The claim is simply that for any *step* in a given calculation, S has a disposition to add finite numbers, carry the digit and move on to the next step. It does not matter how large that calculation is, because *S*'s disposition is focused on the step, not the whole calculation, and formed by learning to perform such steps.

Similarly, no agent has the disposition to sit through a near endless presentation of objects and saying of each one whether it is red or not. It is not unreasonable, however, to think that *S* could have a general disposition to judge whether or not

any *given* object is red.[21] The resulting picture is that, strictly speaking, we do not have the disposition to add any two numbers, but rather have a mechanism that uniquely generates such a disposition for each case through the use of the technique, which in turn requires the use of simpler concepts for which we do have the right kind of disposition. This is what I will mean when I say that mathematical concepts are grounded in techniques.

One might nevertheless ask, given that we have these general disposition about simple cases like 'carrying one', or adding finite numbers, how do these 'ingredient dispositions' combine so that the the second-order equilibrium path covers all cases, beyond our actual answers—without falling prey to my objections to Warren's account, given in Chapter 2? The answer to that question might be best seen by considering a simple example. Suppose that we are considering an example where we do not have any obvious manifest disposition, an example like '135 664 + 37 863 = 173 527'.[22] In order for us to be able to go through the whole process of adding the two numbers here, however, for the purposes of demonstration, suppose that the addition '13 + 9' is in fact not trivial, but of the same kind as the addition problem with large numbers—i.e. one where we do not have any manifest disposition about the sum.

Going through the calculation step by step, we have the manifest, psychologically tractable disposition to judge that 3 + 9 = 12, to carry the one and move on to the next step in the calculation, and finally to judge that 1 + 1 = 2, giving the answer of 22. At each step, however, the game-theoretic structure of the practice provides the correctness conditions. In short, our simple, manifest dispositions combine into an infinite second-order equilibrium path because we also have manifest 'combining' dispositions, like dispositions to carry, that combine into a single stable disposition to judgement at every step, no matter how complex the problem—any number can be 'built up' from simpler manifest dispositions through the technique in a unique way.

21. Consider for instance a neural network which is programmed to recognise pictures of cats (these exist). The network is a finite object, but presumably has the ability to say of any picture whether it contains a cat or not.

If a picture is to large for its memory, it might break them up in to subpictures for which it would still have that ability.

22. By 'manifest disposition', I mean that given a case, I will give an answer right away (of course, only if I am co-operating, not drugged, etc.). Warren's simple dispositions are similar.

It does not matter here that we are discussing '13 + 9', and not '135 664 + 37 863 = 173 527', because every step in *that* calculation is analogous to a step we took in calculating 3 + 9 = 12. It is not that we have some disposition to give a reply at each step by having gone through all previous steps, but rather that for each step, considered locally, we have the disposition to give a reply and move on to the next one. This is exactly analogous to an indeterminate long row of red objects I considered above. It is not the case that anyone has the disposition to go through such a row and say of every object that it is red. However, for each object in the row, considered in isolation from its neighbours and as a result of our training, we do in fact have such a disposition.

The only difference here is that in the case of a calculation, some of these steps in the calculation are essentially 'combining', and that does not mean that they cannot be considered in isolation from all the steps that came before. The end result is that we do in fact have a stable disposition to judgement about every case in an infinite second-order equilibrium path.[23]

### 4.3.2 Objections to communitarian accounts

What then about worries that have been raised against communitarian solutions to the paradox, including the sceptical solution? In Chapter 2, I mentioned what I take to be the main trouble with such solutions, the worry that a communitarian solution results in an account where we cannot speak of the community going wrong.

In general, the idea behind this criticism is, that just like the individual, the community has only computed finitely many sums, and so there is no fact about the community that makes it the case that the next calculation it might make is an instance of the same action as previous ones. By appealing to the community, we've therefore simply moved the problem up a level: there are still going to be different sameness relations from the past to the future for the community as a whole, just like there are for the individual. And hence, if *S*'s meaning is evaluated against the backdrop of the community, then whatever the community says is what goes, and

23. This was essentially Warren's point with the IKEA-model of dispositions. This reply to the objection is very similar to what his account would predict, except it uses the device of the basic constitutive practice to provide correctness conditions at each step.

so there is no room for the notion of the community as a whole making a mistake and no genuine standard of correctness the community is measured against.

However, there is an important difference between how most communitarian accounts are presented and the present account, since correctness is explained as a second-order equilibrium path of a game-theoretic structure here, where each such path picks out a different concept. It follows that locutions such as 'the community calls…', 'the community's dispositions…' and so on, do not have any clear meaning for us—on this account, only agents use words and take part in practices, and they *form* a community. Mistakes are made on the occasion of use and only agents ever use words, not the community. Furthermore, the correct sameness relation from past uses to novel cases in the use of a given term is picked when the structure of the practice is fixed and the agents have acquired their dispositions. There is therefore no question about the community's uses tracking such a relation from past uses to novel cases at all, since one second-order equilibrium path is selected immediately—one that covers all possible cases already.

Hence, if we would all be disposed (i.e. in the sense of having a stable judgement) to call a non-square object 'square', that would simply mean that our word 'square' picked out a different concept than it does now—i.e. not the concept *square*. We wouldn't be deciding that non-squares are square, but rather expressing a different proposition by the sentence '*x* is square' which would either be true or false depending on whether *x* falls under this concept or not. We would be on a different equilibrium path, as it were. The community—i.e. the totality of agents—could conceivably change its dispositions regarding the use of a term, but then the corresponding concept that it refers to would change as well.

It is therefore unclear what it means to make a mistake here: how could the practice pick out the *wrong* sameness relation? Isn't it a primary lesson of the paradox that the idea that one such relation is *sui generis* privileged is misguided to begin with? On this picture, it does indeed depend on us and our judgements to which concepts our words refer, and in this sense we might say that meaning is a judgement-dependent property.[24] But why shouldn't the meaning of *our* words depend to us? It is however not the case that properties *in general* are judgement-dependent on this account: it is up to us what the meaning of the term 'red' means

24. See e.g. Wright 2002 for discussion.

and hence whether or not a particular object falls under the extension of the term 'red' as determined by *that* meaning, but that is far cry from it being up to us whether or not the object *is* red—i.e. falls under the concept *red*.

Again, it is not up to the participants in a basic constitutive practice that some objects are red and some are not—rather it is the *meaning* of the term 'red' that is up to them, represented as a second-order correlated equilibrium of a basic constitutive practice, and this is fixed as soon as their dispositions are fixed, along with the structure of the practice. It is therefore determined in advance, in a certain substantial sense, what the word 'red' means. If we consider a sentence expressing a proposition, for instance, such as 'the letterbox is red', it will be true if and only if the colour of the letterbox is red and the meaning of the word 'red' is *red*. If the dispositions of the agents were to change, then the sentence 'the letterbox is red' would pick out a different proposition with different truth conditions, namely the one where the word 'red' gets its meaning from a different equilibrium. The only thing that varies with the dispositions of the agents is which proposition is picked out by the words they use—the sentence 'the letterbox is red'. But fix an equilibrium path, a meaning for the term, and the agents have no further part to play in whether or not the letterbox is red. *That* depends on the letterbox.

What about a case where everyone is under some kind of illusion? Should we not say that in such a case, the structure of the practice could settle on the intuitively wrong answer, as it were? And then further, that this would be a malign case of the community 'just going'—i.e. one we would want to exclude? I think that a lot depends on how such a case is described. Suppose for instance that every agent belonging to a practice of using the word 'length' is shown two lines on a piece of paper such that the two lines are in fact unequal in length, but by some illusion or another, the agents perceive them as being equal, but we— standing outside the practice and not susceptible to the illusion—would say that they are not equal in length. Further suppose that up to this point, the practice of the agents has been identical to ours, both in terms of actual answers given, but also in that their dispositions are the same as ours as a result of an identical way of coming to acquire those concepts through training.

In this case, we might want to say that the agents are getting it wrong, not that they just have a different meaning of the term 'length'. I think that my account has

no problem delivering that verdict, however, if we suppose that their practice in using the term 'length' is otherwise the same as ours, and hence that their judgements in this case are unstable. For instance, if they also have the practice of measuring length with rulers, laying things on top of each other to see which is longer, etc., then there is an independent way for them to challenge their own dispositions and say that the lines are in fact of equal 'length'—where 'length' is used for *length*. In other words: their practice of using the term 'length' is more complicated than just being based on visual impression and thus settles on the right second-order equilibrium after all, since only one set of dispositions is stable.

It is therefore possible that the actual judgements of every agent in a given case is an unstable one, and that their stable dispositions to judgement settle on a different outcome. This makes room for the possibility of everyone making a mistake, a problem most communitarian accounts struggle with. These cases are structurally similar to horsey-cow cases we discussed in Chapter 2 (Boghossian 1989, pp. 535–536).

If, however, visual impression is the *only* thing they go by, then I do not think there is anything odd in describing them as having a different practice in using the term 'length' and therefore a different meaning, where they aren't getting it wrong.[25] It does not matter if the agents do not get the opportunity to lay the rulers on the line and thus fail to manifest their dispositions about their previous disposition: if their linguistic training with regards to the word 'length' includes rulers and so on, they will have these stable dispositions—and if not, saying that they mean something different by 'length' is the right result.

It is therefore true in a certain sense that there is no room for a mistake for the community as a whole here (thought of as that which sustains a practice), since its role is to pick out one sameness relation from the past to the future as correct, as this selection is in a way arbitrary.[26] There is however room for everyone to make a mistake, and therefore there is a distinction between what the community (thought of as the totality of agents) *thinks* is true and what *is* true.[27] It is not 'up

25. Thanks to Carrie Jenkins and Crispin Wright for the objection this example raises.

26. This doesn't mean that all practices are equivalent. Presumably we have the practice of adding because its a better practice than quadding. In practical terms, adding is useful, quadding is pointless.

27. Another way to make this point is to say that there are no *wrong* basic constitutive practices,

to us' that a particular object *is* red, but rather that it is up to us to which concept the word 'red' refers—the meaning of the term 'red'. That is partially determined by our dispositions to judgement in individual cases—but again, the meaning is fixed when the dispositions of the agents and the structure of the practice itself are fixed, and hence the meaning is fixed in advance of any actual judgement.

### 4.3.3 Self-application of basic constitutive practices

So far in this chapter I've given an account of rule-following and meaning in terms of basic constitutive practices—to mean something by a term '*F*', I said, is to take part in a basic constitutive practice of using '*F*', defined by a second-order correlated equilibrium. On this view, what it *is* to be *F*-ing is determined by this equilibrium and therefore also what it is to correctly use the term '*F*'. Further, for *S* to mean *F* on a particular occasion of utterance is for *S* to intend to *F* and having his utterance lie on the second-order equilibrium of the basic constitutive practice of using the term 'F'.

One might therefore object as follows: If the account is supposed to be fully general, and there is no reason to suppose otherwise, it should also explain the meaning of the word "meaning" (and related terms) as a second-order correlated equilibrium of basic constitutive practice. Now consider the following utterance by *T*:

(1) *S* meant *red* when he said 'red'.

If what *T* said was true, it then follows that *S* meant *red* by his use of 'red' in his original utterance, say

(2) *x* is red.

if and only if *S* intended to say that *x* is red and *S*'s utterance of (2) fell on the second-order equilibrium of using the term 'red'.

However, the original utterance (1) is then correct if and only if *T* intended to say that *S* meant *red* and *T*'s utterance of (1) lay on the second-order equilibrium

but as soon as a practice has been fixed, and hence the meaning of a term, then the account makes room for mistakes, even as made by every member of the community.

of using the term 'means'. But surely *T* didn't mean to say that *S* intended for his utterance to fall on the second-order equilibrium of the practice of using the term 'red'—which is what my account has analysed meaning as. There is a mismatch between what the account analyses meaning as and what itself predicts the meaning of 'meaning' to be. Or, more generally: the word 'red' doesn't mean 'falls on a particular equilibrium path of a basic constitutive practice'—it means *red*.

This is of course correct. I do not think, however, that this objection hits the mark. First of all, there is a distinction between what *constitutes* the meaning of 'red' and what 'red' *means*. My account should be read as an answer to the first question, not the second. The second-order equilibrium path of the basic constitutive practice of using the term 'meaning' is what constitutes the meaning of 'means', but that does not mean that 'mean' means that. The second-order equilibrium of the basic constitutive practice of using 'red' is what explains *that* 'red' means *red* (which sameness relation from the past to the future the term picks out), but it does not follow that therefore 'red' *means* that equilibrium—the right conclusion on this account is still just that 'red' means *red*.

Nevertheless, isn't it possible that these two things might come apart, that (a) *T* correctly says that *S* meant 'red' because *T*'s use of 'means' agrees with the second-order equilibrium of the basic constitutive practice of using the term 'means', but *S*'s original utterance didn't agree with the second-order equilibrium of the basic constitutive practice of using the term 'red'? Or vice versa, that (b) *T* incorrectly says that *S* meant 'red' because *T*'s use of 'means' disagrees with the second-order equilibrium of the basic constitutive practice of using the term 'means', but *S*'s original utterance did agree with the second-order equilibrium of the basic constitutive practice of using the term 'red'? That is to say: isn't it possible that the theory predicts that *T* correctly says that *S* meant *red*, but *S* didn't in fact mean *red* or the other way around?

First of all, this would only be a problem if *S* and *T* both belong to the same basic constitutive practices of using 'red' and 'means', and those practices are the same as our practices—i.e. that we as philosophers reasoning about *S* and *T* are in fact reasoning about *our* terms 'red' and 'means' that refer to the concepts *red* and *meaning*. Further, we are here talking about speaker meaning, not what *S*'s words mean in the language as such, but what *S* meant by his use of 'red'. This

depends on *S*'s intention in making the utterance, and not just the practice itself. With those assumptions in place, I would argue that this is in fact not possible: the account gives the right prediction in this case.

Let's look at (a) first, the case where *T* says that *S* meant red, but *S* didn't in fact mean *red*. I have not developed a full theory of intention (or contentful states more generally) but the basic idea is that *S*'s intention to mean *red* somehow tokens the term 'red' and then the basic constitutive practice of using 'red' of which *S* is a part gives the full content of the intention. A very crude picture (and undoubtedly false) in line with this notion of intention, but nevertheless gives some idea, is that *S* says to himself: 'I'm going to *F* now' and thereby *S* has intended to *F*—what 'F' in *S*'s thought actually refers to then depends on the practice of using *F*. Again, the point is not that this is a good or a fully-fledged account of intention, but just that however such an account would look like in the end, it would understand *S*'s intention to be a matter of *S*'s mental state, in addition to his participation in a basic constitutive practice.

Now, *S* can fail to meet the equilibrium in his use of the term 'red' in two ways: by (i) intending to mean *red* by his use of the term 'red' and not managing to meet the equilibrium (misapplying the term) or by (ii) not intending to use the term 'red' to mean *red* by his use (and there is nothing mysterious about that, perhaps *S* is wilfully using the term 'red' as he normally would use 'green'). In both cases, the account gives the right result. In the former case, *T* is right that *S* didn't mean 'red', just like the account predicts, and in the latter case, given how the agents would have learned to use 'means', *T* would simply be wrong that *S* meant 'red'—i.e. if *T*'s use of the term 'means' the account just doesn't predict that *T* would have used the term correctly in this case. She would simply be wrong about *S*'s intentions.

What about (b), the case where *T*'s use of 'means' misses the equilibrium, but *S*'s original use of 'red' was correct (and *S* did mean *red*)? This case is analogous to (a): We assume that *S* intends to mean *red* and hits the equilibrium. *T* could then miss the equilibrium in two ways: by intending to say that *S* meant 'red' and misapplying the term 'means' or not intending to say that *S* meant 'red' by her use of 'meant'. The account can accommodate *T* misapplying the term as that just means that there was a mismatch between T's intention and the basic constitutive practice of using 'means' that *T* is a part of, and if *T* didn't intend to mean *means* by her

use of 'means' it again gives the right prediction: she shouldn't hit the equilibrium of the practice.

Therefore, the account does give the right result when it is self-applied to the term 'meaning'—at least *prima facie*.

**The basicness of basic constitutive practices** Before moving on, I want to address one further objection. The example I have been relying on through the course of the chapter is that of addition, which I've analysed as a basic constitutive practice. That practice, however, quite obviously depends on the basic constitutive practice of counting, for instance, as well as the basic constitutive practice of carrying, as well as the basic constitutive practice of using other words, perhaps numerals. In what sense then is the practice basic? Am I not explaining language as being composed of many different practices, each of which is completely independent from the others, and could be as they are even if the others did not exist?[28]

It is certainly true that in order to learn addition one has to know how to count and how to carry, both analysed as basic constitutive practices on my account. In that sense, the basic constitutive practice of adding *does* depend on the basic constitutive practice of counting in a meaningful sense: we cannot learn how to add without learning how to count, and our techniques for adding necessarily involve the basic constitutive practice of counting. It is therefore not the case that the two criteria of correctness, that of counting and that of adding are completely independent—even if a statement involving counting is correct because it lies on one equilibrium path, that of its basic constitutive practice, and a statement involving adding is because it lies on another. The paths are in fact dependent on each other, because of how we learn how to add.

That being said, these two practices *do* each have their own equilibrium path which is constitutive of them, and in that sense, they are independent. That can

28. It should be noted that this is a criticism that has been levelled against Wittgenstein by Dummett, who writes that Wittgenstein has a tendency

> ...to regard discourse as split up into a number of distinct islands with no communication between them (statements of natural science, of philosophy, of mathematics, of religion) (Dummett 1959, p. 326)

plainly be seen from Kripke's algorithm example. Suppose for instance, that the method by which we have learned to add two piles of marbles is by first counting $x$ marbles in the first pile and then $y$ marbles in the second, placing the two piles together and counting the resulting larger pile. The result is $x + y$ marbles. It is possible, even if we take the practice of counting for granted, to find a deviant interpretation of this algorithm that does not depend on re-interpreting counting.

For instance, we might interpret it so that for all piles larger than $n$, where $n$ is a number we have never seen before, we *qut* them together, where *quting* is the same as putting for numbers equal or smaller than $n$, but throwing away the first pile after that. The basic constitutive practice of adding does therefore depend on other practices, in the sense that mastery of them is required for mastery of adding, but is independent in the sense that it's own equilibrium path is required to determine if something is an instance of adding or not. Hence, the two practices are independent in the required sense.

## 4.4 Evaluating the account

How then do we stand with the desiderata laid out in the beginning of the chapter? First, lets look at our principle (MC). Recall, this was the principle that for any predicate $F$, an object $x$ is either $F$ or not, independently and in advance of anyone's judgement that it is so. On the account offered here, the correctness conditions for the use of a term is given by a second-order equilibrium path of a basic constitutive practice $P$ of using some term $F$. This equilibrium path represents the meaning of $F$ by constituting what it is to take part in $P$ and by the way $P$ is set up, the meaning of $F$ is fixed as soon as the dispositions to judgement of the agents is fixed, along with the structure of the practice.

This determination of meaning is independent of anyone's judgement that $x$ is $F$, because even though everyone's *dispositions* to judgement play a role in picking out one second-order equilibrium of a basic constitutive practice as the operative one—essentially picking out one relation from past uses to novel cases as metalinguistically correct—this equilibrium just gives the *meaning* of the relevant term, and hence it is the *meaning* of $F$ which depends on everyone's dispositions

to judgement, not *that x* is *F*.[29]

Similarly, our principle (C)—that rule-following has correctness conditions—can be satisfied. As I argued above, that is in fact a quite a trivial matter: if we do not conceive of language as rule-governed, the satisfaction of (C) can be explained by appeal to (MC) which does all the heavy-lifting.

What then about our tripartite distinction between what *S* judges to be the case, what *S*'s community judges to be the case and what really is the case? The distinction between the first and the third is easy: if *S* judges *x* to be *F*, but *x* does not fall in the extension of *F* as given by the second-order equilibrium of the practice of using a term that refers to *F*, then *S* was wrong. The distinction between the second and the third is a bit more complex, but still relatively straight-forward. Recall, the second-order equilibrium of using a term 'F' determines what concept that term refers to, what the extension is and that is determined by the stable dispositions of all the agents taking part in the practice. Since those dispositions are fixed at a given time, it is also fixed at that time what the extension of that term is, for indefinitely many uses in the future, in which case it is possible for every single agent to make a judgement about a given case that is not their stable judgement, and in that way, the whole community can make a mistake. This is not circular, because stable dispositions are not defined in terms of particular judgements that agents make.

Second, what constitutes the meaning of *S*'s utterances is a combination of *S*'s mental state, his intention, and his being embedded in a basic constitutive practice of using the terms that figure in his utterances. In particular, utterances of the form "S meant 'p' " will be true when they meet the second-order equilibrium path of using the term 'means' and here, as explained in the last section, the account gives the right result. We can therefore also account for linguistic mistakes and the problem of error: if *S* intends to say that *p* but does not in fact meet the equilibrium, *S* will have made a linguistic mistake, but not meant something else.

Third, this account does not place place too heavy epistemological burdens on the agent: there is nothing implausible about supposing that *S* has the necessary

29. If I stipulate that all things that have a certain property shall be said to be 'X', I have not thereby made it so that some object *a* is X. *a* is X independently of my stipulation, even if I stipulated the meaning of X. It is the same here.

dispositions to be able to carry out a calculation, for instance, because the necessary dispositions are mediated through a technique we are trained to perform, the correctness conditions of which come from the practice. Similarly, there is nothing implausible to suppose that *S*'s training has resulted in *S* having the disposition to judge for any object whether or not it is red. These judgements explain how *S* can do what a rule or the meaning of a word requires. Guidance, however, is an illusion, since the practice is opaque to the agent.

Finally, we wanted to be able to account for the distinction between merely acting in accord with a rule and be properly following it, as well as accounting for what it is to be justified in making an utterance or apply a rule to a new case. Here, it is the agents intention that makes the difference, where the content of the intention is given by the basic constitutive practice of using the terms which figure in the expression of the intention. As for justification, *S* is justified in his utterances by reference to the practice *S* is taking part in: if *S* writes that 68 + 57 = 125 he is justified because that is what adding is.

### 4.4.1 The Wittgensteinian elements

As I stated in the introduction to this chapter, the solution to the rule-following paradox presented here is not based on exegesis of Wittgenstein's texts and is not Wittgenstein's. Nevertheless, it is intended to provide an explanation what it could mean to say that meaning and rules are practices in a way that might appeal to contemporary analytical philosophers. In particular, I intend to use the account just presented to defend Wittgenstein's radical conventionalism about mathematics in the next part of the thesis.

In this section, however, I'm going to compare the account with Wittgenstein's positive remarks about meaning in *PI* and argue that even if the account is not Wittgenstein's and not in accordance to his general approach to philosophy on the 'no thesis' reading, it nevertheless does solve the problem we set up in Chapter 1 of providing an account of meaning where the key Wittgensteinian notions of practice, form of life, agreement in judgement and so on, do the work the are supposed to do. I will only survey a few remarks, and skip Wittgenstein's remarks on justification and blind rule-following.

First I want to discuss §190:

> It may now be said: "The way the formula is meant determines which steps are to be taken". What is the criterion for they way the formula is meant? It is, for example, the kind of way we always use it, the way we are taught to use it.
>
> We say, for instance, to someone who uses a sign unknown to us: "If by '$x!2$' you mean $x^2$, then you get *this* value for $y$, if you mean $2x$, *that* one."—Now ask yourself: how does one *mean* the one thing or the other by "$x!2$"?
>
> *That* will be how meaning it can determine the steps in advance. (*PI*, §190)

In Chapter 1, we saw that it wasn't clear how to understand this remark. On the account just offered, the *criterion* for what *S*'s meant by '$x!2$' is *S*'s intention *and* the basic constitutive practice of using, in this case, the symbol '!'. However, we can also say that the way the agents in the practice are taught to use the symbol '!' is constitutive of it, as well as the way they always use it (what their stable disposition to use it is). That then, is our answer to Wittgenstein's rhetorical question: it is by being a part of the basic constitutive practice of using the symbol "!" that one means one thing rather than another by "$x!2$".

In §197, Wittgenstein asks how we can grasp the meaning of a word "in a flash". He points out that if we accept meaning mentalism, we are led to think that "the future development must in some way already be present in the act of grasping the use and yet isn't present" (*PI*, §197). He then asks:

> [W]hat kind of super-rigid connection obtains between the act of intending and the thing intended?—Where is the connection effected between the sense of the words "Let's play a game of chess" and all the rules of the game?—Well, in the list of rules of the game, in the teaching of it, in the everyday practice of playing. (*PI*, §197)

On the account offered here, the connection that obtains between an intention and the thing intended is precisely given by the practice of using the terms that refer

to that thing—*S*'s intention of adding gets its content from the basic constitutive practice of using the term '+' of which *S* is a part. That practice relies on how the agents are taught how to use the symbol and how they actually use it.

The very next remark, the famous §198 can be given a similar interpretation. Wittgenstein asks:

> Let me ask this: what has the expression of a rule – say a signpost – got to do with my actions? What sort of connection obtains here?— Well, this one, for example: I have been trained to react in a particular way to this sign, and now I do so react to it.
>
> But with this you have pointed out only a causal connection; only explained how it has come about that we now go by the signpost; not what this following-the-sign really consists in. Not so; I have further indicated that a person goes by a signpost only in so far as there is an established usage, a custom. (*PI*, §198

This suggests a two step picture: first *S* is trained in a particular way to react to sign-posts—in our terminology, acquires dispositions to act—and then there is a regular use, a custom which gives the full explanation, a basic constitutive practice in which *S* is embedded.

The preceding remarks concern training, custom and practice. The last two remarks I want to consider are about our common form of life and agreement in judgements. In §§241–242, Wittgenstein writes:

> "So you are saying that human agreement decides what is true and what is false?" — What is true or false is what human beings say; and it is in their language that human beings agree. This is agreement not in opinions, but rather in form of life.
>
> It is not only agreement in definitions, but also (odd as it may sound) agreement in judgements that is required for communication by means of language. This seems to abolish logic, but does not do so.

On the current account, human agreement in dispositions to judgement in the context of a basic constitutive practice is what determines the reference of our

terms—that '+' refers to *addition*, for instance, rather than *quaddition*. It does not follow from that account that the basic constitutive practice determines *that x* is *F*. In this way, the notion of basic constitutive practice can make sense of what Wittgenstein means by this otherwise quite puzzling remark by explaining how agreement about the particular case, in the sense of agreement in dispositions to judgement, can constitute a practice, while keeping a notion of objective correctness in play– not thereby abolishing logic.[30]

In the next part of the thesis, I will use this account of meaning in order to defend Wittgenstein's radical conventionalism about mathematics. In order to do so, I will read Wittgensteinian locutions such as "it is our practice..." or "we all say that..." as an implicit appeal to basic constitutive practices.

30. See also RFM VI, §40:

> We say that , in order to communicate, people must agree with one another about the meanings of words. But the criterion for this agreement is not just agreement with reference to definitions, e.g. ostensive definitions—but also an agreement in judgements. It is essential for communication that we agree in a large number of judgements.

# Part II

# Wittgenstein's radical conventionalism

# Chapter 5

# Conventionalism, radical and orthodox

Conventionalism about a particular domain is the view that the propositions of that domain owe their truth-value, in some sense or another, to linguistic conventions (see e.g. Quine 1966; Glock 2008; Warren 2015; Topey 2019), or as it is often put, that they are true in virtue of meaning (Glock 2003; Warren 2016).[1] It is also often said that the truths of such a domain, again in some sense, depend wholly on us and how we speak, and not determined by external reality (Dummett 1959, 1993).

Conventionalism about mathematical and logical truth is therefore an attractive position for anti-platonists who do not want to be ontologically committed to such mysterious things as mathematical objects, conceived of independently of and external to our mathematical practices. It is easy to see why: by appealing to a linguistic convention as the ground for mathematical truth, the anti-platonists can offer an alternative explanation of what would otherwise seem an inexplicable difficulty, namely our knowledge of the abstract objects such an explanation posits (see e.g. Benacerraf 1973). And indeed, if mathematical truths are merely linguistic in nature, there wouldn't be anything to explain, since our mathematical

1. In her book on conventionalism, Yemima Ben-Menahem (Ben-Menahem 2006, Chapter 1) argues that conventionalism isn't really about explaining how certain propositions get their the truth-value, but rather the claim that these propositions are mere stipulations, and not truth-apt in the first place. Since such statements are nevertheless *said* to be true, I believe that this is a moot point, and is more about the concept *truth* than it is about any possible conventionalist account.

knowledge is no longer seen as knowledge about objects at all, properly speaking, and fully understandable in light of our mastery of our own language.

In this part of the thesis, I will defend the view that Wittgenstein was a radical conventionalist, as well as argue, notwithstanding some influential arguments to the contrary, that radical conventionalism is a viable view in the philosophy of mathematics. In this chapter, I will introduce Dummett's distinction between radical and orthodox conventionalism (or 'full-bloodied' and 'moderate' as he calls it), as well as canvass two influential arguments against conventionalism, Quine's regress argument and what I will, following Brett Topey (2019), call 'the argument from worldly fact'. The goal of this chapter is to settle on a definition of radical conventionalism and give an impression of why we should *prima facie* be sceptical of that view.

In subsequent chapters, I will argue that Wittgenstein's philosophy of mathematics is indeed a form of radical conventionalism, and that the picture of meaning developed in the first part of the dissertation can be brought to bear in solving the difficulties surrounding the view, including Dummett's influential arguments against it, as well as Severin Schroeder's more recent criticisms, and that consequently, it remains a viable alternative in accounting for mathematical truth.

## 5.1 Quine's regress argument against conventionalism

It is widely held that Quine's 1936 article "Truth by Convention" (1966) decidedly refuted conventionalism about mathematics and logic.[2] Quine's argument proceeds is in two steps. He first considers the idea that every conventionally true statement requires an explicit stipulation: that for a logical statement $\varphi$ to be conventionally true, it needs to have been explicitly stipulated to be true. He quickly points out, however, that since there are infinitely many logical truths, they cannot all be stipulated explicitly.

He next considers the more promising idea that we have stipulated a finite number of axioms to be true and consequently appeal to inference rules that we have likewise conventionally adopted to move from the axioms to further logical truths. These inference rules have, like the axioms, been explicitly stipulated. That

2. See Warren 2016 for numerous examples and discussion.

is to say, we choose our axioms as a starting point, and choose inference rules to get from the axioms to further truths. This, however, leads to a dilemma: as the inference rules we rely on to generate the further logical truths have themselves been explicitly stipulated, we need to rely on some further logical notions to move from the general form of the rule to its particular instances. For instance, in order to draw the conclusion that $q$ from the combination of $p \rightarrow q$ and $p$, we must appeal to the general form of the rule, e.g.

$$\varphi \rightarrow \psi, \varphi \models \psi$$

and some background logical notion to move from that explicitly stipulated general scheme to the particular instance of the rule (universal instantiation perhaps, the rule that if something holds for all, it holds for every particular).[3]

The dilemma then is, that we either land in a regress, explaining those background notions as being explicitly stipulated rules, leading to the same problem again, or we take these background notions for granted, at the cost of having thereby thrown away the conventionalism—as well as the very motivation for the view.

We'd have, as Warren puts it, reduced the conventionalist claim that logic is true by convention to the less exciting claim that "logic is true by conventions *plus logic*" (Warren 2016). In explaining necessity, we have appealed to an unexplained metanecessity, and if that is not conventional, then it is hard to see why necessity should be in the first place (see Dummett 1993, p. 449, for this way of putting the matter). In a nutshell then, Quine's problem is that in order for the orthodox conventionalist position to work, we need to presuppose logic, the very thing to be explained.

It is tempting to take a constitutive or deflationary line here (similar to the one I attributed to Hacker and Baker in Section 1.2 of Chapter 1) that it is simply a part of the meaning of the rule above that its instances are instances of *it*, and hence

3. This version of the problem comes close to Kripke's adoption problem for logic, except here there is no emphasis on the agent doing the reasoning. The point is just that to use an inference rule that has been stipulated to hold, we need other background logical notions.

Kripke's version of the problem is based on unpublished work, but see Padró 2015 for discussion. A recent author who also sees the two problems as being related is Finn 2019.

no problem in explaining how we can move from the general form of the rule to a specific instance. It is just a part of what it means for a rule to be *modus ponens*, for example that those particular instances are instances of that rule. But similar points apply here as they did there: the relation between the explicit formulation of the rule and the rule itself (or its correctness conditions) is contingent, and it is clear that *that* relation cannot be set up by explicit stipulation, without either stipulating infinitely many truths, or a leading to a regress, if general logical principles are supposed to do the trick.

There is of course the obvious possibility of relaxing Quine's demand that explicit definitions need to be used in determining what the convention is, and instead relying on some way of *implicitly* adopting a given convention. Quine does consider this possibility:

> It may be held that we can adopt conventions through behaviour without first announcing them in words; and that we can return and formulate our conventions verbally afterward, if we choose, when a full language is at our disposal. It may be held that the verbal formulation of conventions is no more a prerequisite of the adoption of conventions than the writing of grammar is a prerequisite of speech; that explicit expositions of conventions is merely one of many important uses of a complete language. So conceived, the conventions no longer involve us in a vicious regress. (Quine 1966, p. 98)

Quine thinks, however, that if we were to adopt such a theory, that would mean that conventionalism is no longer doing any explanatory work. Quine writes:

> In dropping the attributes of deliberateness and explicitness from the notion of linguistic convention we risk depriving the latter of any explanatory force and reducing it to an idle label. (ibid., p. 98)

However, Quine does not offer any arguments for why we should consider this implicit form of conventionalism, where nobody in particular stipulates that the convention holds, to be so impotent in explaining mathematical truth, other than claiming that it would not say anything more than the statements of logic and mathematics are *a priori* or merely firmly accepted. That, however, would of course depend on the conventionalist theory on offer.

In Chapter 7, I argue that the picture of rule-following and meaning developed in the first part of the dissertation can serve as a regress stopper against Quine's regress argument and that it does in fact amount to more than the claim that the statements of mathematics are *a priori*. This account, which aims to fill in the gaps in Wittgenstein's appeal to practice and custom, is, I will claim, radical conventionalist in the relevant sense.

## 5.2 The master argument, or the argument from worldly fact

Before moving on, I want to briefly mention a prominent argument against conventionalism that is often taken as being decisive. This argument is variously called the Lewis-Lewy objection (Glock 2003) after early proponents of the argument, C.I. Lewis and C. Lewy (C. I. Lewis 1946; Lewy 1976), the master argument (Warren 2015), or the argument from worldly fact (Topey 2019) and is meant to show that while conventions can determine the meaning of a sentence *s* and so in particular that *s* expresses a certain proposition *p*, they nevertheless cannot make it the case that *p* is true, and hence that conventionalism cannot be right.

The master argument runs as follows: even if it it might be true that it is a convention among English speakers to use the word 'vixen' for all and only female foxes, that tells us only that certain words are applied to certain objects, but what we are really interested in is the fact *that vixens are foxes*—and *that* happens to be true because every individual vixen is a fox, and not because of a linguistic convention. In fact, that would be true even if there were no languages at all. Similarly, the statement "all bachelors are unmarried men" is not true by convention, because the very fact that each bachelor is unmarried is not true by convention—for instance, Kant was not a bachelor because of convention, but because of certain biographical facts peculiar to him. The same goes for any other bachelor, and that is why the statement "all bachelors are unmarried men" is true. Hence conventionalism is false.

Topey summarises the argument thus:

> It's entirely obvious that this sentence [i.e. "all vixens are foxes"] says

> something about the world. In particular, it says something about vixens—namely, that they're foxes. So if this sentence is true by convention, then what it says must be true by convention, which means that our linguistic conventions somehow have the power to make it the case that vixens are in fact foxes, to make the world one way rather than another. But that seems absurd: the conventions of English do make it the case that "All vixens are foxes" says that all vixens are foxes, but it seems clear that whether all vixens are foxes is in no way a matter of the conventions of English or any other language. It's just a matter of what vixens are like. And if that's right, then "All vixens are foxes" can't owe its truth wholly to convention. (Topey 2019, p. 1729)

And here is Ted Sider's version of the argument, worth quoting at some length:

> What could it mean to say that we make logical truths true by convention? Imagine an attempt to legislate truth: "Let every sentence of the form 'If P then P' be true." What would this accomplish? The legislator could be resolving to use the word 'true' in a new way; he could be listing the sentences to which this new term 'true' applies. But this isn't making logic true by convention; it is legislating a new sense of 'true'. (Sider 2003, p. 24)

He then goes on to claim that we cannot make propositions true by simply declaring them to be so:

> There are a number of ways I can cause the proposition that my computer monitor has been thrown out the window to be true. I could throw the monitor out myself, pay or incite someone else to do it, and so on. I cannot, however, cause the proposition to be true simply by pronouncing. I can pronounce until I am blue in the face, and the computer will remain on my desk; my pronouncements do not affect the truth values of statements about computer monitors. (ibid., pp. 24–25)

Statements about conventions, however, are different, Sider claims. He concludes:

> A convention consists of the activities of language users; that is why we can so easily make it the case that conventions exist…Only statements about pronouncements, for example statements about conventions, seem to be made true by our pronouncements. Statements about monitors, or bachelors, or rain, are about a part of the world we cannot affect simply by pronouncement. That it is either raining or not raining is about rain; I cannot affect the world in the matter of rain simply by pronouncement; therefore I cannot make it the case that either it will rain or it will not rain simply by pronouncement. (ibid., p. 24–25)

Boghossian shares the same doubts and asks:

> How can the mere fact that *S* means that *p* make it the case that *S* is true? Doesn't it also have to be the case that *p*? […] How can we make sense of the idea that something is made true by our meaning something by a sentence? (Boghossian 1996, p. 364–365)

Topey argues, however, that it isn't entirely clear that the argument so posed is actually targeting the right view—what conventionalists should really be committed to. Conventionalists are, he says, trying to express the view that sentences of the right sort are not saying anything 'substantive' about what the world is like, that they are 'empty', 'degenerate' or true by definition, and this does not necessarily have to preclude that those statements really are about vixens, rather than just about words. We should, Topey thinks, really be trying to explain what the relevant difference is between sentences such as "All vixens are foxes" on the on hand and "All vixens weigh less than a ton" on the other (Topey 2019, pp. 1731–1732). And here, it seems quite sensible to say that the difference is somehow linguistic.

There is something right about this: it is simply not the case that conventionalists aren't sensitive to the *de re* reading of the statement "all bachelors are unmarried" and think that somehow the marital status of each individual bachelor has been determined by convention—that it was by convention that Kant was unmarried, for example, rather than by circumstance. Conventionalists don't have to be committed to that view, and it is highly dubious that any prominent conventionalists ever did.

Hans-Johan Glock (2003, 2008), for instance, defends conventionalism from a Wittgensteinian point of view and defines a statement *s* to be, in his terms, "broadly analytic" or a conceptual truth, if *A* sincerely denies or rejects *s*, then *A* either didn't understand *s* or is deliberately using *s* in a novel way, or, alternatively, as those statements where the grasp of the meaning of *s* is sufficient for accepting *s*. These definitions of conventionalism are quite close to the explanatory reading offered by Topey, as well as capturing the spirit of something being verbal, degenerate or insignificant about such statements. Furthermore, Glock thinks that what is at stake is finding an explanation of the special epistemological status of analytic statements, not their truth—what needs to be explained is why they are *necessary*, not why they are *true*. That, according to Glock, the conventional bit, not that each vixen is biologically a fox.

The proponents of master argument essentially point out that the truth of the statement "all vixens are foxes" does at least in part depend on the non-linguistic fact that each individual vixen is a fox and that of course human convention cannot create *that* fact. That is of course true, but even if that argument were cogent against these general forms of conventionalism, it does not seem to have much force against conventionalist accounts of mathematics and logic, since the argument would then have to presuppose that mathematical facts exist independently of human practices—e.g. to construe mathematical facts as being like worldly facts in the relevant sense. Of course facts cannot be up to us if by 'fact' we mean something like 'state of affairs, independent of human practices' or 'condition of the external world'—as the property of being a vixen is.

In other words, there is agreement on both sides that a fact such as "Kant is a bachelor" is obviously not a linguistic fact, and the proponent of the master argument tries to use that as leverage to show that the corresponding analytic statement cannot be true because of linguistic convention. Or as Sider put it, "[s]tatements about monitors, or bachelors, or rain, are about a part of the world we cannot affect simply by pronouncement" (Sider 2003, p. 24–25). True enough. But in the case of mathematics, there is no such prior agreement, and this is exactly the picture of mathematical truth that the conventionalist denies—for a conventionalist, the statements of mathematics are simply not descriptions of a mind-independent reality, and so the argument simply presupposes that conventionalism about those

domains is false in order to show that it is so. Glock's gambit, to move the focus from truth to necessity, is therefore unnecessary in these cases. For that reason, however convincing one might find it to be against conventionalism about purportedly analytic statements, I will set the master argument aside.

## 5.3 Dummett on Wittgenstein's radical conventionalism

One of the most influential interpreters of Wittgenstein's philosophy of mathematics is certainly Michael Dummett. Dummett's interpretation was first outlined in a review of the *Remarks on the Foundations of Mathematics*, published in 1959, shortly after the publication of the *Remarks* (Dummett 1959, p. 328), and then decades later in a paper called "Wittgenstein on Necessity: Some Reflections" (Dummett 1993).

Dummett's verdict concerning Wittgenstein's philosophy of mathematics was in most parts negative, not least because of Dummett's reading of the role that the rule-following considerations play in Wittgenstein's philosophy of mathematics. On his reading, Wittgenstein was a radical conventionalist who thought that every step in an inference should be considered as a convention in its own right, unconstrained by what came before. Dummett's motivation for his reading, apart from being reasonably well grounded in the text, is to explain how we can move from finitely many cases in the teaching of a rule to infinitely many applications of that rule—a way to avoid the regress of Quine's argument against explicit forms of conventionalism.

The radical conventionalist therefore agrees that mathematical truth is conventional in the sense that the truths of mathematics and logic depend on us or our language, but in contrast with the orthodox conventionalist, claims that such truths do not depend on adoption of rules, which then operate *independently* of our linguistic practices. The truth of every mathematical statement directly depends on our mathematical practice in its own right, and hence there is no privileged set of stipulated truths. *Every true mathematical statement stands on the same level.*

Dummett offers a barrage of arguments against the view, the totality of which

has been taken as decisive against the view—arguments which I will cover and respond to later in Chapter 7. Dummett also emphasises in his exposition of the view that there is an element of choice present—that for the radical conventionalist, each mathematical truth is *chosen* by us to be true. This is of course highly counter-intuitive, and in the course of the chapter, I will argue that this is not in fact a necessary component of radical conventionalism—without losing either its conventionalist character, nor thereby making the adjective 'radical' no longer apt. That, I will argue, has more do with how to distinguish the view from the more moderate forms of conventionalism, and not an essential component of the view—the essential kernel of radical conventionalism is that every statement is in its own right directly determined by the convention, and not via inference rules that operate independently of our practice.

### 5.3.1 Dummett's definition of radical conventionalism

In this section, I'm going to give Dummett's definition of radical conventionalism and clarify some issues it raises, before I give my own definition of the view. In Chapter 7, I will survey Dummett's arguments against the view, as well as giving my own replies to those arguments, using the solution to the rule-following paradox developed in the first part of the thesis, but for now, I will merely try to explicate the view.

Now, Dummett defines conventionalism in general as the doctrine that

> all necessity is imposed by us not on reality, but upon our language; a statement is necessary by virtue of our having chosen not to count anything as falsifying it. (Dummett 1959, p. 328)

This, it should be noted, is Dummett's definiton of conventionalism in *general*, both in its orthodox and radical varieties, and not merely radical conventionalism.

Dummett's discussion primarily focuses on *necessity* and not *truth* but it is clear from his discussion that he does take the conventionalist position about mathematics to entail that there is nothing more to the truth of a mathematical statement than our linguistic practice, even if the primary goal of the conventionalist is to explain their necessity. There are other things to notice here. First of all, as I mentioned above, Dummett speaks of there being a choice involved: we do not count

anything else than a given necessary truth as being possible because we have *chosen* not to count anything else as being a counter-example. It seems to be a part of this way of explicating the view that we *could*—in some sense—have chosen something else. I discuss the notion of choice involved further below, but for now, this seems particularly problematic for the view, because if we simply choose which statements are correct, the notion of correctness itself seems to have evaporated.

As mentioned in the introduction to this section, Dummett's philosophical motivation in his reading of Wittgenstein as a radical conventionalist is to avoid Quine's regress problem.[4] Here is how Dummett poses the problem:

> It appears that if we adopt the conventions registered by the axioms, together with those registered by the principles of inference, then we must adhere to the way of talking embodied in the theorem; and this necessity must be one imposed upon us, one that we meet with. It cannot itself express the adoption of a convention; the account leaves no room for any further such convention. (ibid., p. 329)

The orthodox conventionalist, Dummett thinks, relies on the notion of logical consequence to explain how we move from the axioms to the theorems, but logical consequence is what (among other things) the theory itself is purporting to explain and is on this view something that is imposed upon us from without, independent of our actual practice, and not a mere convention.

According to Dummett, Wittgenstein avoids this unfortunate consequence by going in for radical conventionalism (although Dummett calls it "full-bloodied conventionalism"). Dummett's first pass at explaining this view is as the view that all necessary statements are so because they have been adopted directly as conventions. For Wittgenstein, Dummett writes,

> the logical necessity of any statement is always the direct expression of a linguistic convention. (ibid., p. 329)

This, one might say, indicates that what mostly separates the orthodox conventionalist from the radical is that for the latter, the source of the necessity of every

4. It might be noted here that Putnam (1979, p. 424) dubs this the 'Quine-Wittgenstein objection' to orthodox conventionalism, as he sees Wittgenstein's considerations of rule-following as structurally similar.

mathematical statement is the same: there are no privileged statements from which further truths derive their necessity, everything is on one level, and each truth is determined *directly* by the convention, but not a *consequence* of that convention, conceived of independently of our practice. Another way of putting the same point, is that if conventionalism relies in some sense on agreement (however that is specified) being the source or ground of truth, then the radical conventionalist position is that agreement *about the particular case*—not general rules or stipulations—is that ground.

Dummett, however, expresses this by saying that each truth is directly *stipulated* to be true, since for Dummett, the conventionalist picture generally sees necessity as a result of having chosen certain statements to be exempt from counter-examples. Hence, he thinks of radical conventionalism as a view where each individual truth is so chosen:

> That a given statement is necessary consists always in our having expressly decided to treat that very statement as unassailable; it cannot rest on our having adopted certain other conventions which are found to involve our treating it so. (Dummett 1959, p. 329)

Thus, the necessity of any given necessary statement, no matter how complex, is always explained by Dummett's Wittgenstein as an explicit choice of not accepting anything as counting against it, but not as a necessary consequence of our having adopted any other more basic conventions.

Dummett takes the following example as illustrative. The first criterion we accept in saying whether or not there are *n* things that fall under a certain concept is the procedure of counting. But if we find that there are five boys and seven girls in a room, we say that there are twelve children, without finding it necessary to count all of them as a group. That we are justified in this is not because the fact that "5 + 7 = 12" is implicit in the practice of counting, as the moderate conventionalist would say, but because we have chosen to adopt *this* as a criterion for this sum independently of any counting. As Dummett points out, if these criteria really are independent of each other, it is not inconceivable that they may clash with each other either—otherwise they would not be truly independent. The necessity of "5 + 7 = 12" is therefore to be explained as privileging it over any other statement if

such a clash occurs, for example, if we count up the five boys and seven girls and find that they are eleven, we say that we *must* have miscounted, but not that we had thereby discovered that this particular arithmetical statement was after all false.

The important thing about this example for Dummett, is that according to our normal intuitions, which Dummett himself shares, as soon as we have laid down our rules of inference or in this case arithmetical rules, we do not have any "further active part to play" but as Wittgenstein's considerations about rules make clear (on Dummett's interpretation) we are free to accept or reject a proof at any step, as there is

> nothing in our formulation of the axioms and of the rules of inference, and nothing in our minds when we accepted these before the proof was given, which of itself shows whether we shall accept the proof or not; an hence there is nothing that *forces* us to accept the proof. (ibid., p. 330)

If, Dummett concludes, we accept the proof, "we confer necessity on the theorem proved; we "put it in the archives" and will count nothing as telling against it" (ibid., p. 330).

In Dummett's later paper, he does not take himself to deviate from his previous interpretation, but considers an objection to the view presented here, in order to clarify it. Suppose, for instance, that we have accepted a proof that a cylinder intersects a plane in an ellipse. According to radical conventionalism, we have therefore acquired a new criterion for applying the term "ellipse" which we might appeal to in certain cases to say that some figure, while not looking like an ellipse, must nonetheless be one. It could then be objected that there could therefore never arise a circumstance in which a counter-example might lead us to doubt our theorem, and subsequently discover a mistake in the proof. This is ruled out, because the correctness of the proof is simply taken to be our acceptance of it as a proof, and since that very acceptance is supposed to make us rule out any counter-example *a priori*, we could never doubt our own proof. But, the objection goes, this has in fact happened many times in the history of mathematics, and so Wittgenstein is merely confusing necessity with certainty (Dummett 1993, p. 447).

But, Dummett argues, this is not what we should take the view to entail. We should modify it so that what we have already proved should only be taken to be *provisionally* compelling, admitting perhaps of a counter-proof or an empirical counter-example. What is important, Dummett says, is that we do not introduce the *ideal* into the account—saying that what is necessary is what the ideally competent mathematician would call necessary, for example, as that would entail that there are external standards of judgement that exist independently of us and our own mathematical practice, and *that*, Dummett thinks, is what the radical conventionalist denies.

Dummett contrasts this again with the modified or orthodox conventionalist (which he calls "restrained conventionalism" in this later article). On that view, again, all necessary truth is derived from linguistic stipulations we have made and are trained to observe. Some necessary truths, like "there are seven days in a week" are, Dummett says, direct "subjects" of such stipulations and learning them and treating them as true is required for learning the meaning of the word "week". Other truths, however, are more remote consequences of such stipulations or conventions. This prompted the objection that this form of conventionalism leaves it unexplained how *those* consequences are necessary.

Radical conventionalism, on the other hand, as Dummett puts it, treats "*every* necessary truth as the direct expression of a linguistic convention" (Dummett 1993, p. 447). That, Dummett points out, does not erase the distinction between basic truths which cannot be given any further justification than perhaps their own meaning, as is the case with the truth "there are seven days in a week", and truths that are consequences of such truths and admit of proof. The real difference between the radical conventionalist and the moderate is that those consequences, i.e. the truths that are not basic in this sense, are so because it is also our convention to accept them as such, or as Dummett puts it:

> [T]here is no sense in which they would be consequences whether we recognized them as such or not. (ibid., p. 447)

For the radical conventionalist, there can be no criterion external to our practice as for what counts as such a consequence. That does not mean that we cannot regard some truths as more basic, e.g. axioms, nor that we cannot revise our practice with

criteria *internal* to it.

It should be emphasised here, how much Dummett's own reply to this objection jettisons the emphasis on decision as integral to radical conventionalism. If decision at each step was the defining element of the view, and not the individual conventionality of each truth, the repudiation of the ideal, or the rejection of external standards and criteria, then Dummett's reply simply would not work, as the correct step is on that reading defined as what we have decided, and hence we would have *no* criteria at all, not even internal to the practice, to later doubt our choice. In his later paper, Dummett himself does therefore not treat decision as integral to radical conventionalism.

In his later paper, Dummett cites Putnam's description of his reading of Wittgenstein with approval. It runs as follows, and is worth quoting at some length:

> Wittgenstein was a conventionalist who held not just that some finite set of meaning postulates is true by convention, but that whenever we accept what we call a 'proof' in logic or mathematics, an act of decision is involved: a decision to accept the proof. This decision, on Dummett's reading, is never forced on us by some prior thing called 'the concepts' or the meaning of the words; even given these as they have previously been specified, it is still up to us, whether we shall accept the proof as valid deployment of those concepts or not. The decision to accept the proof is a further meaning stipulation: the 'theorems of mathematics and logic' that we actually prove and accept are not just consequences of conventions, but individually conventional. (Putnam 1979, p. 424 as cited by Dummett 1993)

In this description of the view, we see a lot of the same language as in Dummett's own version. Not only, in Putnam's view, is radical conventionalism a view where each statement is a direct expression of a convention where every necessary truth stands on the same level, but it is also the case that accepting a proof is an 'act of decision' to 'accept' the proof which us never forced on 'us', and so on.

**'Decision' as metaphor** What role should these phrases, 'decision', 'acceptance', 'us', etc., play in our definition of radical conventionalism? Dummett is not entirely

clear what he means when he says that 'we decide at each step how to go on' or that we 'chose which statements are necessary', but given what he says about conventionalism more generally, he seems to have something like explicit stipulation on Quine's model in mind, that there is in some way, either by the individual rule-follower or the community as a whole, an *act* of decision or acceptance involved. We somehow actually *choose* how to go on at each step. Naturally, this has been held to be a wildly implausible view—of course we do not *choose* how to go on.

In Chapter 7, I will discuss this further, but for now, I will only say that these locutions should not be taken literally, but rather be seen as metaphors for what is really important about the view, namely that mathematical statements are not responsible to anything external to our mathematical practices and that the truth of each mathematical statement is directly determined by the convention and not mediated through rules which operate independently of our practice. There is nothing like a decision going on in the strict sense of the word, much less an 'act of decision'. We should rather understand 'decision' as being used impersonally (even though, of course, Dummett and Putnam do seem take it literally) to signify the fact that nothing outside of human conventions and practices are what determines the outcome in each case—that in some substantial sense those truths are up to us, that there are no standards of judgement external to our practice.

In fact, Wittgenstein himself seems to be somewhat uncomfortable with the talk of decision and often qualifies it or takes it back. The most prominent example, is of course already from §186:

> It would *almost* be more correct to say, not that an intuition was needed at every stage, but that a new decision was needed at every stage" (emphasis mine).

Wittgenstein makes similar statements in *RFM* and in the *Lectures on the Foundations of Mathematics*. Many commentators have therefore argued that this one word—'almost'—shows that Dummett's reading can't be right—of course Wittgenstein didn't think that we just decide how to go on and whatever we so decide is correct. I argued against that reading in Chapter 1.

However, insisting that decision is a necessary element in defining radical conventionalism, I believe, a mistake, as this is not is what essential about Dummett's

reading, even by his own lights—namely the contrast between orthodox conventionalism and its radical counterpart. Dummett understands the former kind of conventionalism as a position where some sentences are privileged as being directly adopted by speakers as being true or made true by stipulation, but further truths are derived from these as consequences which are imposed on us from without.

In the latter case, that of radical conventionalism, every sentence has the same status from the point of view of the convention. So, as we saw above, Dummett believes that there's an element of choice in the more moderate forms of conventionalism as well, but only for the most basic truths—and as such, his notion of choice in the case of radical conventionalism really boils down to there being nothing external that constrains our practice, or as Putnam put it, for the radical conventionalist, each step

> ...is never forced on us by some prior thing called 'the concepts' or the meaning of the words; even given these as they have previously been specified, it is still up to us.

And hence, if one does not think that this element of choice is necessary for moderate conventionalism, it can hardly be said to be necessary for the radical variety.

Furthermore, even Dummett's own general definition of conventionalism about some domain more generally does not depend on the notion of choice, but rather on the idea that there is nothing external to our language that determines the truth of the propositions of that domain: "all necessity is imposed by us not on reality, but upon our language", as he puts it. Even for Dummett himself then, the idea that conventionalism in general requires deliberate choice is neither here nor there, and there is no reason to follow him in defining radical conventionalism in that way either. And indeed, in his later article, he does not emphasise that aspect, and given his argument against the objection that we could never reform our practice, implicitly seems to have abandoned it, while stressing the idea that for the radical conventionalist, nothing external to our practice can ever constrain it.

Yemima Ben-Menahem makes a similar point in her book on conventionalism (Ben-Menahem 2006), emphasising that the notion of free choice should not be understood from the point of view of individual rule-followers, but rather as being a claim about the logical properties of the rule itself:

> Wittgenstein does not compare the *experience* of rule following to that of making a free choice. But Dummett does not ascribe such a comparison to Wittgenstein. The point is a logical one, pertaining to a rule's logical power to determine its applications. On reflection, however, this defence of Dummett is not fully satisfactory, for the difference between the phenomenology of obeying a rule and that of making a free choice is reflected in grammar. If it follows from Dummett's solution that this grammatical difference is not a *real* difference (because 'ultimately' the rule's force is illusory), then the problem remains. (Ben-Menahem 2006, p. 258f)

The problem she is alluding to here is precisely the problem of how to give a coherent account of what Wittgenstein means when he says that to follow a rule is a practice, a problem addressed in the first part of the dissertation.[5] It follows, if Ben-Menahem is right, that if that account is suitably conventionalist, then we can solve this problem and retain Wittgenstein's conventionalism. I argue that this is possible in Chapter 7.

Finally, in his paper cited with approval by Dummett, Putnam offers a different definiton of radical conventionalism as the claim that "the truth of the theorems as well as the axioms arises from us" (Putnam 1979, p. 432). My proposal amounts to accepting this definiton, rather than the emphasis on decision.

### 5.3.2 A revised definition of radical conventionalism

In the preceding chapters, I've defended the view that the meaning of words is given by what I called basic constitutive practices. Even if one does not accept that account of meaning, it should be relatively uncontroversial that for Wittgenstein, meaning is grounded in practice and custom—however that is spelled out. For that reason, I will define radical conventionalism about mathematics to be the view that

5. The problem, as put by her, is:

> The dilemma remains: if Wittgenstein's view is indeed a form of conventionalism, we must find in his writings a response to the rule-following paradox that reestablishes a suitable bond between a rule and its applications, so that the notion of convention can replace the traditional notion(s) of necessity. (Ben-Menahem 2006, p. 257)

(RCM) true statements of mathematics are true in virtue of—or grounded in—our mathematical practices such that each truth is directly determined by those practices, without any external constraint or criterion of correctness.

This definition is intended to contrast with orthodox conventionalism and the term "directly" is meant to do most of the heavy lifting in indicating this—by this I mean that *what* follows from *what* is also explained by the means of the convention, or perhaps rather, our practice, without any external criterion. That is to say, radical conventionalism does not take the notion of logical consequence for granted, but explains that in terms of our practices as well, instead of seeing it as being imposed on us from without. Another way of putting this point is that for the radical conventionalist, rule-following is also explained by appealing to our conventions or practices.

(RMC) is the view that I will defend in subsequent chapters, but I will mostly refer to it simply as *radical conventionalism*. This definition does not mention that these truths are linguistic truths, but in light of our background view of meaning which I will use to defend the view, there is no loss of generality here: our basic constitutive practices of using mathematical symbols and phrases is what gives their meaning (that is, correctness conditions) and hence explain the truth of the mathematical statements in which they figure. In essence, the radical conventionalist equates mathematical truth with correctness in our mathematical practice. In turn, I will explain correctness in a mathematical practice by appealing to basic constitutive practices, and thus use the the solution developed in Part I to defend radical conventionalism.

In the next chapter, I will offer a close reading of the relevant aspects of Wittgenstein's philosophy of mathematics in the *Lectures*. The core idea that Wittgenstein seems to defend there is that it is our agreement about the *particular case* which is constitutive of the concepts we use, and I will argue that this is indeed the essence of radical conventionalism, as each truth is thereby directly determined by our practice.

# Chapter 6

# Wittgenstein's philosophy of mathematics in the *Lectures*

In this chapter, I will give quite a detailed overview of Wittgenstein's philosophy of mathematics in the *Lectures* as it pertains to the nature of mathematical truth and related matters, and argue that he is best seen as a radical conventionalist—i.e. holding the view that true mathematical statements are linguistic truths, grounded in our mathematical practices, such that each such truth is a direct expression of those practices, without any external constraint or criterion of correctness.[1]

It is important to make a distinction at the outset, however, between radical conventionalism about mathematics, i.e. the position discussed and defined in the previous chapter and just outlined, and a view about rule-following and meaning that might accompany it, i.e. a similar view about rules and meaning in general, for example the view about meaning defended in Chapter 4. The former does not follow from the latter, although both might be described as 'radical conventionalism'

I will argue that (a) for Wittgenstein, the meaning of a word or a rule is constituted by our agreement *about each particular case*, without any prior commitments or external constraints, and (b) that in the case of mathematical statements, Wittgen-

1. There are of course other interesting topics in the *Lectures*, e.g. Wittgenstein's views on contradiction and inconsistency, as well as his claim that mathematical statements are norms of description. To keep things focused, however, I will not cover these things here.

stein's view is that *nothing* outside of our language and mathematical practice determines their truth. If we combine these two views, it follows that mathematical statements are true in virtue of our mathematical practice such that each truth is determined directly by that practice, and not a consequence of any prior commitments or external constraints—which is of course nothing but radical conventionalism about mathematics.

The main aim of this chapter is to argue that Wittgenstein did indeed hold both views just outlined, and hence that he was a radical conventionalist about mathematics. In the next, and last, chapter, I will argue that we can explicate the otherwise puzzling claim that meaning is constituted by agreement about the particular case by re-interpreting it through the solution to the rule-following paradox offered in Chapter 4—i.e. I will argue that we can explain what it means for our agreement about the particular case to constitute the meaning of our words by identifying this agreement with our stable dispositions to judgement that determine a second-order equilibrium path of basic constitutive practice.

This chapter will be in two parts, the first on rule-following and understanding in the *Lectures*, where I argue that Wittgenstein saw agreement about the particular case as constitutive of the correctness conditions of the concept involved, and the second on Wittgenstein's conventionalism about mathematics, where I argue that he thinks that the ground for mathematical truth is our mathematical practice and language.

## 6.1 Wittgenstein's discussion of rule-following in the *Lectures*

The arguments and examples that make up the rule-following considerations of §§185–242 of the *Philosophical Investigations* are quite prominent in the *Lectures* and Wittgenstein returns to them time and time again. They do not, however, as they do in the *PI*, receive a systematic treatment nor does Wittgenstein explicitly say that to follow a rule is a practice or a custom. In this section, I will discuss Wittgenstein's remarks on rule-following in the *Lectures* and argue that Wittgenstein's position is that training and human agreement about the particular case

play a *constitutive* role in the determination of meaning—what we all *would* do, Wittgenstein seems to be saying, in a particular case is a part of what determines the correctness of that same case.[2]

In Lecture II, Wittgenstein first brings up arguments that are reminiscent of his discussion in the *Philosophical Investigations*. The first use he makes of these arguments is to demonstrate that understanding is not an occurrent mental state:

> What is a momentary act of understanding?
>
> Suppose that I write down a row of numbers
>
> 1 4 9 16
>
> and say, "What series is this?" Lewy [Casimir Lewy, one of Wittgenstein's students in attendance][3] suddenly answers, "Now I know!"—It came to him in a flash what series it is. (*LFM* II, 27)

Wittgenstein then goes on to argue that even if the correct formula, namely '$y = x^2$', had come into Lewy's mind when he suddenly expressed understanding, that in itself would not have guaranteed that Lewy would have continued correctly, nor that he had in fact understood, since Lewy might always understand *that* formula in different ways.[4] Wittgenstein then makes the point that appealing to the notion of 'the same'—that insisting that Lewy would just apply the formula that occurred to him in the same way as was being done in the examples—gets us no further (see discussion in Chapter 1 and Chapter 2).

---

2. I should mention that the famous §201 of the *Philosophical Investigations* supports this reading, even though it has not been traditionally read that way. Anscombe's original translation reads:

> What this shews is that there is a way of grasping a rule which is *not* an *interpretation*, but which is exhibited in what we call "obeying the rule" and "going against it" in actual cases.

while the most recent translation has, instead of "actual cases", the more confusing and less readable "from case to case of application".

The original, "von Fall zu Fall", might be more idiomatically rendered as "on a case-by-case basis".

3. This is the same C. Lewy who was among the first to advance the argument from worldly fact against conventionalism. See the last chapter.

4. This argument is of course reminiscent of *PI,* §138.

Wittgenstein next supposes that we teach Lewy to square numbers by giving him a rule and examples that range from 1 to 1,000,000. He points out that a natural reaction to what he has just said is to reply by saying that we can then *never* know if Lewy has understood what we were trying to teach him: no matter how often Lewy demonstrates his competence, there are other rules that fit with what he has done. But that, Wittgenstein says, is not the real problem:

> But the real difficulty is, how do you know that yourself understand a symbol? Can you really know that you know how to square numbers? Can you prophecy how you'll square tomorrow?—I know about myself just what I know about him; namely that I have certain rules, that I have worked certain examples, that I have certain mental images, etc., etc. But if so, can I ever know if I have understood? (*LFM* II, p. 27–28)

The point is not, however, scepticism about meaning and understanding. The lesson Wittgenstein draws from is discussion is, yet again, that meaning and understanding are not kinds of *mental acts* "which [anticipate] all future steps before we make them" (*LFM* II, p. 28).[5] Nevertheless, Wittgenstein asks,

> Should one then say that if I write $y = x^2$, where $x$ is to take all the intergers, that it is not determined what is to happen at any particular point? (*LFM* II, p. 28)

The ensuing discussion parallels the discussion of §189 and later in the *Philosophical Investigations* in many ways (see Section 1.1.2 for an overview of those remarks).

There, as well as here, Wittgenstein seems to only allow two senses of the word "determine", namely that it may mean (a) that "people trained in a certain way generally go on writing down a certain series" and that "they all act in the same way when confronted with this formula and asked to write down its series" (*LFM* II, p. 28) or (b) as a statement about the mathematical form of the formula, i.e. to contrast formulas like $y = x^2$ and $y = x^z$—the latter of which determines infinitely many series, each depending on the value of $z$.

However, after being pressured on the point by one of his students, he says:

5. The emphasis here is on the acts, not the anticipation. See Chapter 1 for discussion on Wittgenstein's repudiation of meaning mentalism.

> "Does the formula '$y = x^2$' determine what is to happen at the 100th step?"
>
> This may mean "Is there any rule about it?"—Suppose I gave you the training below 100. Do I mind what you do at 100? Perhaps not. We might say, "Below 100, you must do so-and-so. But from 100 on, you can do anything." This would be a different mathematics.
>
> If it means, "Do most people after being taught to square numbers up to 100, do so-and-so when they get to 100?", it is a completely different question. The former is about the operations of mathematics but the latter is about people's behaviour. (*LFM* II, p. 29)

There seems to be a tension in what Wittgenstein says here. At first, he seems to be allowing that there is a third way to understand the question, showing that he's not simply forgetting this possibility, but then again only seems to offer these two alternatives: either the question is about people's behaviour (that there is a rule about it seems to be reducible to people's training) or it is a mathematical question. After a brief interlude about the role of intuition in rule-following, he says:

> But a man is only said to know by intuition that $25 \times 25 = 625$ if 625 is in fact the result which we all get by calculation. But a man is said to know $1 + 1 = 2$ not because two is in fact the result which we can reach by calculation—for what sort of calculation should we use?—but because he says with the rest of us that $1 + 1 = 2$.
>
> The real point is that whether or he knows it or not is simply a question of whether he does it as we taught him; it is not a question of intuition at all. (*LFM* II, p. 30)

Wittgenstein certainly seems to be saying here that the way that a person is trained in using symbols, on the one hand, and agreement *about the particular case*, albeit about a very simple case, on the other, is constitutive of the correctness of each step.[6] He then goes on to say that following '|, ||' by '|||' or going from '1 to 2 to 3, etc.' is "more like an act of decision than of intuition."

6. This is contrasted with agreement about general principles—agreement that we are supposed to follow these rules, e.g. the rules of multiplication. For example, Severin Schroeder reads Wittgenstein in this way. He writes:

That does certainly seem like evidence in favour of Dummett's reading of Wittgenstein's radical conventionalism, rather than mine, namely that there is an act of decision involved, but Wittgenstein then immediately clarifies:

> But to say "It's a decision" won't help [so much] as: "We all do it the same way". (*LFM* II, p. 31)[7]

Again, the idea that 'we all do it the same way' seems to be how Wittgenstein wants to solve the problem of rule-following—that somehow the very fact that we all do it, or presumably would do it, in the same way is constitutive of the correctness conditions of the rule.

At the start of the next lecture, Wittgenstein summarises his discussion of the matter in a way that makes his emphasis on training come out quite clearly:

> We saw that the word "determine" can be used in two different ways. One can ask "Does my pointing determine him to go in a certain direction" and mean by that question either "Will he (or most people) go in a certain direction when I point?" or "Is one trained in such a way that, when I point, it is correct to go in a certain direction and incorrect to go in other directions?" (*LFM* III, p. 32)

Again, it seems that Wittgenstein wants to reduce (or perhaps rather, explain) objectivity in rule-following by appealing to what most people will do or how they are trained, and that this is in fact all there is to the matter.

Wittgenstein returns to the question of rule-following again in Lecture VI in the context of a discussion about the notions 'same', 'analogous' and 'similar'. Here, Wittgenstein gives perhaps the most explicit formulation of the view I'm attributing to him in the *Lectures*, namely that training and human agreement about a particular case is constitutive of correctness in the use of symbols. After describing

---

> Empirically speaking, there is no social agreement on this particular sum, there is social agreement only on the general principles of multiplication (Schroeder 2017a, p. 95)

See Chapter 7 for further discussion.

7. It should be noted that there is some doubt that this section was reconstructed correctly from the notes of the students. I'm relying on the *content* having been reported correctly, which seems plausible given that Wittgenstein says similar things often.

a case where one shows a partner how to do certain movements and then asking them to do the same thing, and explaining how they might then misunderstand any such instruction, but typically don't, Wittgenstein says, and it is worth quoting at some length:

> Similarly one can show a child how to multiply 24 by 37, and 52 by 96, and then say to it, "Now multiply 113 by 44 analogously." The child may then do one of many things. If he can't justify his action, we should go through it again and again, until we converted him to doing the same as us. *The only criterion for his multiplying 113 by 44 in a way analogous to the examples is his doing it in the way in which all of us, who have been trained to do it the same as us, would do it.* If we find that he cannot be trained to do it the same as us, then we give him up as hopeless and say he is a lunatic. (*LFM* VI, p. 58. Emphasis mine.)

Here, Wittgenstein clearly and explicitly says that the only criterion for whether or not a rule has been followed is whether or not the rule-follower in question does it in the same way as those that have received the same way would do it. Wittgenstein's use of the counterfactual is also significant, since he doesn't seem to be saying that this agreement needs to be manifested in every case to do the work it is meant to do.

Later in the lecture, Wittgenstein discusses this picture in relation to mathematical proofs and explicitly uses the word "convention" to describe his view:

> Mathematical conviction might be put in the form, 'I recognize this as analogous to that'. But here "recognize" is used not as in "I recognize him as Lewy" but as in "I recognize him as superior to myself". He indicates his acceptance of a convention. (*LFM* VI, p. 63)

And given what Wittgenstein had just said, that the only criterion for doing the same thing in a particular case is to what others do, it seems reasonable to suppose that view he is advancing is indeed what I described as the one of the natural components of the radical conventionalist position—that without agreement about a particular case, there is no right or wrong, correct or incorrect.

In Lecture X, Wittgenstein essentially makes this point in a discussion on the difference between experiment and calculation. He has made the point that in

calculations there is such a thing as right and wrong, while in experiments there is not, and then considers a case where multiplication is being invented and that so far only numbers below 100 have been multiplied together. He then considers a particular case, namely 123 × 489, and suggests that we might ask someone to do the same thing for these two numbers as we did for the numbers below 100. This, Wittgenstein says, would be an experiment, but one whose result we might adopt as a calculation. He explains:

> What does that mean? Well, suppose 90 per cent do it all one way. I say, "This is now going to be the right result." The experiment was to show what the most natural way is—which way most of them go. Now everybody is taught to do it—and *now* there is a right and wrong. Before there was not. (*LFM* X, 94)

Here, Wittgenstein is explicitly considering a particular case and says that *if* everyone is taught to do it that way, then that is the correct way—and further that the correctness is constituted by that agreement after training ("before there was not").

The alternative reading, that Wittgenstein is in fact not considering agreement about a particular case as constitutive in these examples, but the general way how subjects of the experiment handle numbers greater than 100 is not plausible, because Wittgenstein even more explicitly considers another particular case a little later in his discussion, namely the case of '12 × 12 = 144':

> Russell said, "It is possible that we have always made a mistake in saying 12 × 12 = 144." But what would it be like to make a mistake? Would we not say, "This is what we do when we perform the process which we call 'multiplication'. 144 is what we call 'the right result' "?
>
> Russell goes on to say, "So it is only probable that 12 × 12 = 144." But this means nothing. If we had all of us always calculated 12×12 = 143, then that would be correct—that would be the technique. (*LFM* X, p. 97)

Here, Wittgenstein is clear that *if* our practice were such that we would all say that 12 × 12 = 143, a particular case different from what we actually do, then that

would be correct. Here, the claim, as indicated in the introduction to this chapter, is not that we decide what the outcome is, but rather that our agreement about the particular case determines to *which* concept the symbol '+' refers to in the first place—if we all say $12 \times 12 = 143$ it is to the concept which gives that as the correct answer, and if we all say $12 \times 12 = 144$, then that concept is *addition*. We are not simply agreeing *that* $12 \times 12 = 144$, but rather it is our agreement about this particular case that constitutes the fact that our practice is addition, but not quaddition.

Our agreement about the particular case, Wittgenstein seems to be saying, is constitutive of what our practice is, that if we would all do it one way, then *that* would be our practice, rather than a mistake in another practice—our actual practice where $12 \times 12 = 144$. That is to say, if we let $\times_{144}$ be the multiplication function and $\times_{143}$ be a function that agrees with the multiplication function in every place, except let $12 \times_{143} 12 = 143$, then the claim is that our agreement determines which of these we are in fact referring to, and so if we would agree that $12 \times 12 = 143$, then that would thereby show that the function we refer to by '×' is not $\times_{144}$, but $\times_{143}$. That would, as Wittgenstein says, then be our practice.[8]

This is tantamount to claiming that nothing outside our mathematical practice, not even prior committments, can serve as a criterion of correctness in individual cases, since otherwise it *would* be conceivable that we could make such a mistake as Russell suggests—and that is of course just an expression of radical conventionalism. The point is worth emphasising: For Wittgenstein, it is inconceivable that we would all make a mistake in a relatively basic case like $12 \times 12 = 143$, since if we would all judge that $12 \times 12 = 143$, that would be our practice, and hence correct. The practice is here the only criterion of correctness, as per the definition of radical conventionalism.

But how do we move from facts about all of us agreeing to a certain result—itself an empirical fact—to the corresponding mathematical proposition according to Wittgenstein? On this matter, he says, and it is again instructive to quote at some length:

8. Finding the right way to spell out the notion of agreement is the subject of the next and last chapter. I will argue that by identifying 'agreement' with a second-order equilibrium path in a basic constitutive practice, we can give an account that preserves objectivity in our mathematical practice, as well as avoids certain other problems.

> It has been said: "It's a question of general consensus." There is something true in this. Only—what is it that we agree to? Do we agree to the mathematical proposition, or do we agree in *getting* this result? These are entirely different. [...] Mathematical truth isn't established by their all agreeing that it's true—as if they were witnesses of it. Because they all agree in what they do, we lay it down as a rule, and put it in the archives. Not until we do that have we got to mathematics. One of the main reasons for adopting this as a standard, is that it's the natural way to do it, the natural way to go—for all these people. (*LFM* XI, 107.)

This difference in kinds of consensus, that something is a matter of opinion or "agreement in witnessing" on the one hand, and agreement in action on the other, is one that Wittgenstein often emphasises (see below and e.g. *PI* §241). It is not clear, however, what this difference is, and Wittgenstein often struggles to articulate it. In the next chapter, I will argue that the solution offered in Chapter 4 can elucidate what that difference is.

There is a puzzle here, however. Despite this emphasis on practice and agreement about the particular case being the ground of correctness when using symbols (and indeed being the sole criterion of mathematical truth, as we will see later), Wittgenstein does not deny that it is an objective fact that certain things follow from certain rules, and both in the *Lectures*, as well as in RFM and the *Investigations*, he goes to great lengths to prevent that interpretation of his position. For instance, he explicitly rejects the idea that we are somehow free to stipulate what the result of an individual calculation is:

> We have learned the rules of multiplication, but we have not learned the result of each multiplication. It is absurd to say that we invent $136 \times 51 = 6936$; we find that this is the result. (*LFM* X, 101)

However, if our agreement about a particular case is constitutive of the correctness of that case, how is that not tantamount to claiming that we *invent* that $136 \times 51 = 6936$? If mathematical truth is a matter of "putting something in the archives", as Wittgenstein puts it, how can a case nobody has ever seen before be similarly treated? Presumably, we haven't put $136 \times 51 = 6936$ in the archives until we

have calculated it, and so isn't Wittgenstein's claim somehow circular? A similar objection to radical conventionalism is due to Severin Schroeder (2017a) and I will address it in more detail in the next, and last, chapter.

After a discussion on what might happen if disagreement about multiplication were to arise (where Wittgenstein says that the notion of two different groups seeing different analogies is not clear), Wittgenstein declares:

> The fact is that we all multiply the same way—that actually there are no difficulties about multiplication. (*LFM* XI, 109)

In a later lecture, Wittgenstein made a similar point, emphasising the different kinds of agreement. He first points out that people tend to react in the same way after having gone through the same training:

> If you have learned a technique of language, and I point to this coat and say to you, "The tailors now call this colour 'Boo' ", then you will buy me a coat of this colour, fetch one, etc. The point is that one only has to point to something and say: "This is so-and-so", and everyone who has been through a certain preliminary training will react in the same way. If I just say "This is called 'Boo' " you might not know what I mean; but in fact you would all of you automatically follow certain rules.

He then asks:

> Ought we to say that you would follow the *right* rules?—that you would know the meaning of "boo"? No, clearly not. For which meaning? Are there not 10,000 meanings which "boo" might now have?—It sounds as if your learning how to use it were different from knowing its meaning. *But the point is that we all make the SAME use of it*. To know its meaning is to use it in the same way as other people do. "In the right way" means nothing. (*LFM* XIX, 183)

Here it might seem that Wittgenstein is expressing meaning scepticism, since he says there is no such thing as "the right way"—but that reading would be too quick, since immediately after having made this remark, he goes on to say that this is the

same for continuing the series of cardinal numbers (and presumably any other series) and here the criterion of correctness *is* doing it in the same way as everyone else:

> Is there a criterion for the continuation—for a right and a wrong way—except that we do in fact continue them in that way, apart from a few cranks who can be neglected? (*LFM* XIX, 183)

His previous denial of there being a correct way should therefore be understood as the radical conventionalist claim that there is no such criterion outside of our practice. That reading is supported by his conclusion:

> This has often been said before. And it has often been put in the form of an assertion that the truths of logic are determined by the consensus of opinions. Is this what I'm saying? No. There is no *opinion* at all; it is not a question of opinion. They are determined by a consensus of action: a consensus of doing the same thing, reacting in the same way. There is a consensus, but it is not a consensus of opinion. We all act the same way, walk the same way, count the same way.

This last remark might at first glance be read as a rejection of conventionalism, however, rather than an affirmation—since Wittgenstein says that the truths of logic are *not* determined by a consensus of opinions. However, mathematical statements are true, we are told, because of the other kind of consensus, the consensus of *action* (the same distinction Wittgenstein had been trying to make before). If so, it seems again that Wittgenstein is saying that agreement about the particular case *is* constitutive of the correctness conditions of rule-following and meaning after all, again an expression of the radical conventionalist position.

Before moving on to the next section, where I look at Wittgenstein's remarks about mathematical statements more specifically, I want to look at one remark that makes it clear that Wittgenstein does not regard rule-following as a matter of decision, at least not in the individual case. In Lecture XXV, Wittgenstein is discussing the notion of rule-following in proof, and how self-evidence is not a criterion for correctness. He then states:

> Suppose I tell you to multiply 418 by 563. Do you *decide* how to apply the rule for multiplication? No; you just multiply. Probably no rule at all would come into your head. And if one did, no other rule for the application of the first rule would come into your head. It is not a decision; nor is it an intuition. (*LFM* XXV, p. 238)

If rule-following is not a matter of decision in the individual case, as Wittgenstein seems to be denying here, it might seem natural to say that it is a decision in some other sense, perhaps a collective decision. However, as I argued in the last chapter, this is not a necessary component of radical conventionalism, and in the next chapter I will argue that the necessary notion of agreement does not require an act of decision.

## 6.2 Mathematics and correspondance to reality

In the last section, I argued that Wittgenstein's position on rule-following in the *Lectures* is one where our agreement about the particular case is constitutive of the correctness of that case—if, Wittgenstein seems to be saying, we all agree that $12 \times 12 = 143$, then that is correct. In the next chapter, I will offer interpret this through the solution to the rule-following paradox offered in Chapter 4 and identify our agreement with a second-order equilibrium path in a basic constitutive practice.

In this section, however, I will look at Wittgenstein's discussion of mathematics and its correspondance to reality. The core claim I will defend is that for Wittgenstein, mathematical statements are not responsible to an external and mind-independent reality, but rather truths of language.

The statement can be put in various different ways. It can be read as saying that mathematical statements are not propositions at all, and thus are not descriptions of anything, much less an external mind-independent reality (Bangu 2012b) or a rejection of the view that the ultimate ground for the correctness of our mathematical statements is a mathematical reality, conceived of as independent of our mathematical practice and language (Gerrard 1991). The formulation I wish to focus on here is the latter, although I think it is also quite certain that Wittgenstein

held a version of the former (see e.g. *AWL*, 152, *RFM* I, §144, *RFM* I, app. III, §§1–4). In either case, it is first and foremost a rejection of mathematical platonism, the view that mathematical reality is the ground of mathematical truth.

The guiding metaphor Wittgenstein uses for his rejection of a mind-independent reality serving as the ground for mathematical truth is is claim that the "mathematician is an inventor, not a discoverer" (*RFM* I, §168, *RFM* I, app. II, §2, see also *LFM* II, 22). By this, or so I read him, he does not mean that mathematical statements are not objective, but rather that their ultimate justification lies in our mathematical practices, not in a reality that stands outside them and "adjudicates the correctness" (Gerrard 1991, p. 126) of those practices. He does not think that locutions such as "a reality corresponds to our mathematical propositions" are necessarily false, but that a "wrong picture goes with them" (*LFM* XIV, 141) and unless we provide an explanation of this correspondance, we have simply said something meaningless.

Lecture XXV contains perhaps the clearest expression of this aspect of Wittgenstein's philosophy of mathematics. He says:

> Suppose we said first, "mathematical propositions can be true or false". The only clear thing about this would be that we affirm some mathematical propositions and deny others. If we then translate the words "It is true..." by "A reality corresponds to..."—then to say that a reality corresponds to them would say only that we affirm some mathematical propositions and deny others. We also affirm and deny propositions about physical objects. [...] If that is all that is meant by saying that a reality corresponds to a mathematical propositions, it would come to saying nothing at all, a mere truism: if we leave out the question of how it corresponds, or in what sense it corresponds. (*LFM*, XXV, 239)

Wittgenstein then goes on to say that the words of our language have various different uses and that if we forget where the expression "a reality corresponds to" is "really at home" (*LFM* XXV, 240) we are liable to be misled (see also *PI*, §116). He ends this train of thought by saying:

> What is "reality"? We think of "reality" as something we can *point* to.

> It is *this*, *that* (*LFM* XXV, 240).

It is presumably that picture, on Wittgenstein's view, that invites the comparison with empirical descriptions, saying that a reality corresponds to the statement would in fact be apt.[9]

Wittgenstein next brings the focus back on to the interpretation of the locution "a reality corresponds to..." which he had declared harmless, namely that mathematics is objective. He says:

> Or to say this [that mathematical statements correspond to a reality] may mean: these propositions are *responsible* to a reality. That is, you cannot just say anything in mathematics, because there is the reality. This comes from saying that propositions of physics are responsible to that apparatus—you can't say any damned thing.
>
> It is almost like saying, "Mathematical propositions don't correspond to *moods*; you can't say one thing now and one thing then. Or again: "Please don't think of mathematics as something vague that goes on in the mind." [...] And if you oppose this you are inclined to say "a reality corresponds". (*LFM* XXV, 240)

Wittgenstein then offers two different ways we could in fact spell out this phrase "a reality corresponds..." and give it content (*LFM* XXV, 241). The first way is the one that he calls "mathematical responsibility" which is how certain mathematical propositions but not others can be derived from our axioms by our inference rules—where we might say that a theorem is responsible to the axioms from which it was derived. This way of being responsible to a reality is internal to mathematics itself and amounts to the claim that mathematics is objective, in Wittgenstein's sense above.

---

9. It is quite common among commentators to claim that Wittgenstein is not rejecting mathematical platonism as being false, but only a truism or a confusion. Given Wittgenstein's opposition in the *Lectures* to the idea that mathematical statements are not descriptions of an external reality, the preceding discussion should show that this squeamishness is unwarranted: Wittgenstein's claim is that if we do not provide an explanation of how mathematical statements correspond to reality we have said something confused or truistic, but he is happy to rule out some such explanations as being wrong (and hence false), for instance, ones that posit an external reality.

The second way is the way that the whole system of axioms and inference rules, not merely individual propositions, can be said to be responsible to something. Of the second way, Wittgenstein claims that this is in a certain sense arbitrary, subject to constraints. The constraint that concerns us here is that if we use a word in a particular way, we are inclined to use it in certain ways in future cases and here some such ways to proceed are 'unnatural'.[10] Here Wittgenstein is quite (and perhaps uncharacteristically) explicit:

> Suppose I said, "If you give different logical laws, you are giving the words the wrong meaning." This sounds absurd. What is the wrong meaning? Can a meaning be wrong? There's only one thing that can be wrong with the meaning of a word and that is that it is unnatural. (*LFM* XXV, 243)

What does Wittgenstein mean by "unnatural"? He takes two examples. The first is of us using the words 'red' and 'green' as we use them now, but also going on and describing things as being "reddish-green"—we do not, Wittgenstein seems to be saying, know how to use such a proposition, given the meanings of the words 'red' and 'green' (but we might *set up* a use for it, perhaps, as Wittgenstein suggests, using that word to describe a certain "iridescent black" that otherwise red and green leaves can be in the fall (*LFM* XXV, 243)). The other example concerns the way we count:

> And it is unnatural for us, though not for everyone in the world, to count "one, two, three, four, five, many." We just don't go on in that way. (*LFM* XXV, 243)

This notion of "naturalness" is one that comes up again and again in the *Lectures*, and one that is especially important for Wittgenstein's radical conventionalism

10. The other constraint concerns contradictions. Wittgenstein says:

> If we allow contradictions in such a way that we accept that *anything* follows, then we no longer get a calculus, or we'd get a useless thing resembling a calculus. (*LFM* XXV, 243)

The implication seems to be, first of all, that if we do not allow anything to follow from a contradiction, then we can still have a useful calculus, and, secondly, that it is conceivable that we do so, and if we did, then that would be in some way problematic. (This short quote already shows that Wittgenstein wasn't quite as cavalier about contradictions as many have supposed.)

about mathematics. We, he seems to be saying, simply find some ways of following a rule in a new case more natural than others and that nothing more needs to be said about the correctness of certain ways of continuing, that those ways are in fact how we all proceed and they are therefore constitutive of that very practice (see e.g. *RFM* I, §116).

Early in the next lecture, Wittgenstein brings up the topic of correspondance to reality again, where he elaborates on the two constraints, bringing usefulness to the foreground. He makes a distinction between how we might say that a reality corresponds to a true empirical statement, such as 'it rains' and how we might say the same of individual words, such as 'rain' (*LFM* XXVI, 247). The first, he seems to be saying, is the same as the statement being true or assertable: we might say that a reality corresponds to 'it rains' if it is true that it is raining.

The latter sort, Wittgenstein says, is quite different and amounts to asserting that the word has meaning and showing how such a word corresponds to reality is to give the word a meaning. Accordingly, propositions of that sort, i.e. "this is green" or "'rain' means this" are, as Wittgenstein puts it, "sentence[s] of grammar"—sentences used to explain the use of the word in question (*LFM* XXVI, 248) and set up a meaning for them in our practice. In the case of 'green' and 'rain', Wittgenstein says, we might point e.g. to green things (or out the window when it is raining, supposedly) and thereby explain the meaning of the words 'green' and 'rain'.[11]

The situation is quite different, Wittgenstein thinks, with words like 'two' or 'perhaps'. There are things, Wittgenstein thinks, that we can point to in these cases to explain the meaning of such words, we might for instance raise two fingers, point to them and say 'this is two' (and that would, Wittgenstein thinks, be a perfectly adequate definition of 'two'). However, if we then said 'a reality corresponds to 'two'', Wittgenstein says, it would not be clear at all what we mean:

> The point is this. We can explain the use of the words "two", "three", and so on. But if we were asked to explain what the reality is which corresponds to "two", we should not know what to say.—This? [*He*

11. An utterance like "there are six people in this room", however, is not a sentence of grammar, but an affirmation of a proposition, despite 'six' appearing in it.

> *indicated the two fingers.*] But isn't it also six, or four? (LFM, XXVI, p. 249)

The appeal of the idea that a reality corresponds to such words (and here Wittgenstein includes not just 'two' and 'perhaps', but also 'and', 'or' and 'plus') Wittgenstein attributes to the fact that we have a use for such words in our practice. He concludes by claiming that the way mathematical statements correspond to reality is like how such words correspond to reality, and that this is just a matter of our linguistic practice:

> What I want to say is this. If one talks of the reality corresponding to a proposition of mathematics or of logic, it is like speaking of a reality corresponding to these *words*—"two" or "perhaps"—more than it is like talking of a reality corresponding to the *sentence* "It rains". Because the structure of a "true" mathematical proposition or a "true" logical proposition is entirely defined in language: it doesn't depend on any external fact at all.[12] (*LFM* XXVI, p. 249)

He then goes on to claim that to say that a reality corresponds to a mathematical statement like "2 + 2 = 4" is like saying that a reality corresponds to 'two' in that there is nothing that we can point to directly to give it meaning, and *that* in turn is like saying that a reality corresponds to a rule, which again,

> would come to saying: "It is a useful rule, *most* useful—we couldn't do without it for a thousand reasons, not just *one*." (*LFM* XXVI, 249)

Wittgenstein then concludes:

> You might say: Mathematical and logical propositions are still *preparations* for use of language—almost as definitions are. It's all a put-up job. It can all be done on a blackboard. We just look at the signs and

12. It is quite common among commentators to claim that Wittgenstein is not rejecting mathematical platonism as being false, but only a truism or a confusion. The preceding discussion should show that this squeamishness is unwarranted: Wittgenstein's claim is that if we do not *provide an explanation* of how mathematical statements correspond to reality we have said something confused or truistic, but he is happy to rule out *some* such explanations as being wrong (and hence false), for instance, ones that posit an external reality.

> go on here; we never go outside the blackboard.—The correspondance of mathematical propositions to reality is like the correspondance of negation to reality. It is all entirely *independent* of the other correspondance with reality, the correspondence of "it rains" It's like the correspondance of a word to something used in an ostensive definition. (*LFM*, XXVI, p. 249)

The idea seems to be, that when we utter statements like "this is two" or "'rain' means this" in the contexts Wittgenstein is considering, we are not describing anything, but setting up the meaning of those words. These statements are rule-like in that they tell us how to use the terms in question on future occasions—kind of like constitutive rules for our own practice of using the terms, e.g. the trivial constitutive rules of Chapter 4. These rules are justified by how well they fit into our linguistic practice, which in turn has a practical point.[13]

At the end of the lecture, when discussing certain geometrical statements, he puts it thus:

> What you are saying [when uttering a geometrical statement] is not an experiential proposition at all, though it sounds like one; it is a rule. That rule is made important and justified by reality—by a lot of most important things. (*LFM* XXV, 246)

He then goes on to describe how certain experiential facts stand behind our mathematical statement that $21 \times 14 = 294$, for instance that we can arrange matches in 21 rows of 14 and count them, that we all get this result when we do, and would agree that if we didn't that a match had been added or vanished.

This picture is undoubtedly a conventionalist one, and given that Wittgenstein emphasises our agreement about the particular case as being constitutive of the correctness of that case, arguably a radical conventionalist one.

13. This aspect of Wittgenstein's philosophy of mathematics is undoubtedly interesting and important, however, it cannot be covered here. For a more detailed discussion, see Hacker and Baker 2009, chapter VII and Schroeder 2014.

There is also a quite an extensive literature on Wittgenstein's claim that mathematical propositions are hardened empirical statements. See e.g. Fogelin 1995; Steiner 2000, 2009; Bangu 2012a. For a discussion on the connection between such hardening and work in experiential psychology, see Bangu 2012c.

In Lecture XIV, Wittgenstein discusses an explicit case of mathematical truth on this picture, that of Goldbach's conjecture—the unproven conjecture that every even integer greater than 2 can be expressed as the sum of two primes. There, he claims that to believe that Goldbach's conjecture is true without having a proof of it is to believe that we will find it most natural to extend our mathematics in that way:

> Suppose someone had a hunch that "every even number greater than 6 is the sum of two primes". If you have a hunch it will come out right, you have a hunch that the mathematical system will be extended this way—that is, that it will be best or most natural to extend the system in such a way that *this* will be said to be right. (*LFM* XIV, p. 137)

If we agree that our agreement about a particular case is constitutive of our use of our symbols and that there is nothing to the truth of mathematical statement other than our language and mathematical practices, it then follows that our agreement about a particular case like Goldbach's conjecture is constitutive of the truth of that statement: our practice *could* be extended so that the conjecture is true and it *could* be extended so that it is false. And that is precisely what Wittgenstein says next:

> Suppose someone said: "What you, Wittgenstein, say comes to saying we could *also* extend arithmetic in such a way as to prove this is not so, or to make it a primitive proposition." I'd say: certainly. (*LFM* XIV, p. 137)

He continues:

> Because of course you haven't made this extension. The road is not yet actually built. You could if you wished assume it isn't so. You would get into an awful mess. (*LFM* XIV, p. 137)

A little later, he states that having a hunch that the conjecture is true is "a hunch that people will find it the *only* way of proceeding" (*LFM* p. 138). It would not be an understatement to say that this view is extremely counter-intuitive and does not fit well with what we think we are doing when we are doing mathematics—the

phenomenology of proof.[14] Furthermore, we tend to think that if something is a mathematical fact, it is because it could not have been otherwise—if Goldbach's conjecture is true, it is because that is how the structure of the natural numbers *really* is, independently of us.

Wittgenstein is of course aware of this:

> You might say, "Wittgenstein, this is bosh. For if the system will be extended in such a way, it must be *capable* of being extended in such a way." If this is so, then the person who has a hunch that Goldbach's theorem is correct has a hunch about the possibilites of extension of the present system—that is, he believes something about the essence, the nature, of the system, something mathematical about it. (*LFM* XIV, p. 137)

Wittgenstein attributes this view to Turing, and says to him:

> If you say, "The mere fact that a proof could be found is a fact about the mathematical world", you're comparing the mathematician to a man who has found out something about a realm of entities, the physics of mathematical entities. If you say, "You can this way or that way", you say there is no physics about mathematics. (*LFM* XIV, p. 138)

Here, Wittgenstein explicitly equates the view that mathematical facts are independent of us with the stronger platonist view that mathematical statements are *about* mathematical entities.

But even if Wittgenstein does think that our mathematics can be extended either way, in some sense at least, he does not think that we are completely unconstrained in what we do:

> The mathematical proposition says: The road goes here. Why we should build a certain road isn't because mathematics says that the road goes there—because the road isn't built until mathematics says it goes there. What determines it is partly practical considerations and *partly analogies in the present system of mathematics*.

14. I will discuss this case in more detail in the next, and last, chapter.

> But the fact that a proof of the theorem is *possible* may seem to be a mathematical fact—not a fact of convenience etc. (*LFM* XIV, p. 139. Emphasis mine.)

What are these constraints then? It seems that the view Wittgenstein is advancing is that even though our practice *could* be extended such that the conjecture is true and extended in such a way that is false, we will in practice only find one way of extending it natural or practical. But isn't that a contradiction? If our practice can only be extended in one way, in what sense *could* it be extended in two ways? For now, I will only say that if we accept that whichever we do in fact say, then that is constitutive of the concepts involved—i.e. if we say that Goldbach's conjecture is true, that is one possible practice of using the symbols that figure in the statement of the conjecture and if we say it is false, that is another—then there is room to make a distinction between these two different kinds of possibility: if we would find it natural to extend the system in another way, then that would be our way, and hence correct, but as a matter of fact, we only do find one way natural, and in so far as what we *would do* is fixed in advance, it is also fixed in advance whether or not Goldbach's conjecture is true or not. That, however, only amounts to saying: believing that the theorem is true without proof is a belief about our concepts, and if we agree that our concepts are fixed by our agreement about particular cases, a belief about what we all would agree. Or as Wittgenstein puts it, a belief about what we all find natural to say.

This idea that mathematics is extended by "what we find natural to say" is of course quite vague. In the next chapter, I will defend it by identifying it with 'stable dispositions to judgement'. Here, however, I want to respond to one objection that is natural to raise at this point. Suppose that we find it natural to say that Goldbach's conjecture is true, and hence that every even integer greater than 2 can be expressed as the sum of two primes. Now further suppose that we find a counter-example, a concrete and decidable example of an even integer which cannot be expressed as the sum of two primes. In this case, we would presumably no longer find it natural to say that Goldbach's conjecture is true, unless we would also find it natural to throw out addition and multiplication, by denying that the counter-example is one.

First of all, if we bracket the notion of 'naturalness', which is admittedly quite

unclear at this point, Wittgenstein's claim seems to revolve around how we extend *proofs*—rather than the theorems themselves. The idea is not that we simply find it natural to say that Goldbach's conjecture is true, *sui generis*, as it were, and therefore it is true. It is rather that we piecemeal extend a proof, each little step of which is correct because of our agreement that it is so, taken because we find it natural to extend it that way, ending with the statement of the theorem. The same point would apply to the purported counter-example. If it is decidable that it is so, then there is some calculation that we can perform to show that it is indeed a counter-example. In this calculation, each step is also correct because of our agreement about that step, which is taken because we find it natural.

Therefore, if the scenario described would happen, that would only show that arithmetic is inconsistent, which is the result we should expect, since no philosophical account of arithmetic can protect it from inconsistency, not even a realist one. The point is that we are heavily constrained by our mathematical practice in what we could possibly find natural, and it is not the case on this picture that everything goes. If we would find it natural to say, after each little step in our proof, that Goldbach's conjecture is true, and natural to say, after each little step in our calculations, that there is a counter-example to that theorem, and after thoroughly reviewing our proofs and calculations we still find them in order, then that would mean that arithmetic is inconsistent.[15]

In the next lecture, Wittgenstein brings the subject up again, with a different example, namely the fact that one cannot mate with two bishops and a king:

> Doesn't this "we can't mate with two bishops" rest on a mathematical fact? We might say: It is a question of mathematical possibility. The question is "Is it possible?"—not whether anyone will ever try it. Isn't there such a thing as mathematical possibility?

Wittgenstein's argument against this way of putting the matter is to refer yet again

15. This is not a point about inconsistency not being important or that we can ignore it. The point is just that under the circumstances described, our practice would in fact be inconsistent. That would be the case even if we re-described it under any other account of mathematical truth. If, however, we would immediately make it a primitive proposition that Goldbach's conjecture is true, as Wittgenstein does seem to allow, then that would just mean that arithmetic would be immediately inconsistent, or as he put it "we would get into an awful mess".

to his rule-following considerations. He starts, however, by comparing the idea of there being mathematical facts with the idea of there being a mathematical reality:

> Frege, who was a great thinker, said that although it is said in Euclid that a straight line *can* be drawn between any two points, in fact the line already exists even if no one has drawn it. The idea is that there is a realm of geometry in which the geometrical entities exist. What in the ordinary world we call a possibility is in the geometrical world a reality. In Euclidian heaven two points are already connected. (*LFM* XV, p. 145)

He then says:

> We multiply 25 × 25 and get 625. But in the mathematical realm 25 × 25 is already 625.—The immediate [objection] is: then it's also 624, or 623, or any damn thing—for any mathematical system you like.—If there is a line drawn there between two points, there are 1000 lines between the points—because in a different geometry it would be different. [...] You might say, "I want to go into a world where a straight line really does connect two points."—Yes, but there is an infinity of those. And an infinity of consequences follow, etc.
>
> You never get beyond what you've decided yourself; you can always go on in innumerable different ways.

The idea is, I believe, that if we want to say that the truth of Goldbach's conjecture or that it is not possible to mate with two bishops is determined by facts determined independently of our practice, e.g. by mathematical reality, there is still the question of *what* mathematical concepts we are in fact using—presumably, there is a number-like structure such that Goldbach's conjecture is true and a number-like structure such that it is false, just as there is a chess-like game that deviates from actual chess in just the right way so that "mating" with two bishops is possible—say if it is played directly under the Eiffel Tower at a full moon, to paraphrase one of Kripke's sceptical hypotheses (this would, naturally, be a mate–like concept which agrees with our *mate* in all the right ways, except...)

Our practice up to now underdetermines which of these concepts we are using, and hence underdetermines which structure is the actual one, and which game we are in fact playing. Hence, appealing to mathematical facts that exist independently of our practice does not get us any further, or so Wittgenstein seems to be arguing, since both sets of facts would therefore exist in the relevant sense, and our practice underdetermines which one is in play.

Instead, Wittgenstein suggest that our "*language* is the shadowy reality" that grounds mathematical facts (*LFM* XV, p. 146) and that the truth of the statement "It is impossible to mate with two bishops and a king" is a conceptual truth based on the fact that "there is nothing I will show you here that you will ever call "mating with two pawns"." (*LFM* XV, p. 148).[16] The idea, again, seems to be that our agreement about particular cases in using the chess pieces is constitutive of the correctness conditions of following the rules that govern them, and given how we have learned chess and other facts about us, there will simply never be anything that we will ever call mating with two pawns—it is unnatural for *us* to extend the relevant chess concepts to new and new cases in that way—and since it is constitutive of our use of the pieces how we extend their use, the purported mathematical fact is grounded in our agreement, not an an external reality or other constraints.

Likewise, the truth (or falsity) of Goldbach's conjecture is based on the empirical fact about us that we will only find one way of extending our mathematical practice natural—any other way will not be what we would do, and hence not constitutive of the concepts we are using when expressing the conjecture. Or so Wittgenstein seems to be saying.

The picture of mathematical truth that arises from this discussion is a very counter-intuitive one—and has been sharply criticised and and universally rejected, in large part because it is seen to have consequences that are simply impossible to square with the objectivity of mathematics.[17] In the next chapter, I will use

16. Wittgenstein switched examples in the middle of his discussion. First he spoke of bishops, and then of pawns.

17. In one of the lectures, Wittgenstein addresses these concerns by Turing:

> Turing doesn't object to anything I say. He agrees with every word. He objects to the idea he thinks underlies it. He thinks we're undermining mathematics, introducing Bolshevism into mathematics. But not at all. (*LFM* VI, p. 67)

the solution to the rule-following paradox developed in Part I to argue that it is indeed possible to be a radical conventionalist without rejecting mathematical objectivity—that radical conventionalism about mathematics is after all not so radical and that it is a viable view in the philosophy of mathematics.

# Chapter 7

# Defending radical conventionalism

As outlined in the last chapter, Wittgenstein's position in the *Lectures* was (a) that our agreement about particular cases in the application of a rule or the use of a word is constitutive of the correctness conditions of the concept being used, and hence the correctness of that particular case—i.e. if we agree (in some sense or another) that the proposition '12 × 12 = 143' is true, then the concept referred to by the symbol '×' in that proposition *is* the concept which is such that '12 × 12 = 143', and not actual multiplication where 12 × 12 = 144.[1] When we extend our practice into new cases, hitherto unseen or uncalculated, what we find 'natural' (and hence what we *would* all agree to) is what determines correctness. In the case of mathematical statements, (b) Wittgenstein's view is that nothing outside of our mathematical practices and language grounds the truth of mathematical statements. The combination of these two positions entail radical conventionalism about mathematics.

Thus put, the resulting account is vulnerable to the arguments against community solutions outlined in Chapter 2. In particular, the crucial notions of 'agreement' and 'naturalness' are unclear and ill-defined, and seem to undermine objec-

1. By 'actual multiplication', I mean the concept that I, the author, am referring to when using the symbol '+', and presumably you, the reader, as well.

The point can perhaps also be made in this way. Let $\times_{144}$ be the multiplication function and $\times_{143}$ be a function that agrees with the multiplication function in every place, except $12 \times_{143} 12 = 143$. The claim is that our agreement determines which of these we are in fact referring to, and so if we would agree that 12 × 12 = 143, then that would thereby show that what we are referring to by '×' is not $\times_{144}$, but $\times_{143}$.

tivity in mathematical practice. In this chapter, I will argue that Wittgenstein's position can be re-interpreted through the solution to the rule-following paradox outlined in Chapter 4 and these notions can thus be given a more precise meaning. Since this account of rule-following and meaning is able to avoid the standard objections to community solutions to the paradox, as well as making room for the notion of objectivity in rule-following, in particular the possibility of a mistake, a radical conventionalist account of mathematics that takes it as a foundation is able to as well.

I will then examine a number of powerful arguments against radical conventionalism, in the first instance due to Michael Dummett and Severin Schroeder. I finally address an argument by Putnam, the so-called consistency objection. I argue that given the account just outlined, these arguments fail and that radical conventionalism remains a viable view in the philosophy of mathematics.

## 7.1 Radical conventionalism through basic constitutive practices

In Chapter 4, I identified the concept that a term $t$ refers to with a second-order equilibrium path of a basic constitutive practice of using $t$. On that account, agents acquire stable dispositions to judgement about the particular case through their linguistic training, e.g. dispositions to judge whether or not patch of colour is 'red', and subsequently, the meaning of the term 'red' is given (or constituted) by the game-theoretic structure of their basic constitutive practice of using the term 'red'—i.e. the second order equilibrium path through all possible uses of the term, thereby establishing that it refers to the concept *red*.

This practice is constitutive, because the second-order equilibrium path defines what it *is* to be taking part in the very practice of using the term 'red'—on this view, something is red if only if the term 'red' applies to it, and that is the case if and only if referring to that patch as red lies on the second-order equilibrium path of using the term 'red'. In other words, the correctness conditions of rule-following and meaning are constituted by basic constitutive practices.

Furthermore, this determination is made via our stable dispositions to judge-

ment about each case, and so our agreement about each case—in the sense of having the same stable dispositions to judgement that come together in the game-theoretic structure of the practice—is constitutive of the concept we are using to judge that case: if our dispositions were different about that case, then the equilibrium path of the practice would be different, and hence the concept being used. In that way, our agreement in dispositions about each case is constitutive of the practice itself.

Since the structure of the practice and the dispositions of the agents are all given at a particular time, the second-order equilibrium path settles the correctness conditions of every subsequent use of the symbol '+' in advance, as the inputs already cover every possible case. That is to say, if we assume that the stable dispositions to judgement of all the agents in a given basic constitutive practice are fixed, then a second-order equilibrium path that covers every possible case has *eo ipso* also been fixed. Therefore, there is no question of the community 'deciding' that a particular patch is red, in the sense that the critics of community solutions have used the term, even if the totality of the agents' stable dispositions to judgement that it is red is constitutive of the concept *red*.

The point, I should emphasise, is not that the dispositions of the agents track the concept *red*—conceived of independently of the practice of using the term—but rather that the stable dispositions to judgement that the agents do in fact have regarding the use of the term 'red' define which concept the term 'red' refers to for them. Crucially, because the notion of 'stable judgement' isn't defined with reference to the outcome of those judgements, there is room for the whole community to be wrong about a particular case—there is in other words no analogue of the 'problem of error' for the community on this account, and hence there is a robust notion of objectivity in play. We can all judge that a thing is red, without stably judging that a thing is red. For a more detailed discussion and replies to objections, I refer to Chapter 4.

**Applying the account to Wittgenstein's position in the Lectures** In the previous chapter, I argued that for Wittgenstein, correctness is constituted by our 'agreement about the particular case'—that if we would all say that $12 \times 12 = 143$, then that would be our practice and hence correct. Our agreement, it seems

Wittgenstein is saying, is constitutive of what our practice in fact is. Naturalness, or what we would all find natural to say, on the other hand, was meant to determine how we extend our usage to new and hitherto unseen cases.

We can make these notions more precise by identifying 'agreement about a particular case' with a point on a second-order equilibrium path and 'naturalness' with our stable dispositions to judgement—with that which we are in fact stably disposed to judge about a particular unseen case, which in the good case lies somewhere on a second-order equilibrium path. This agreement is determined by what we are disposed to judge, as well as the structure of the practice. As on Wittgenstein's picture, training and a common form of life, in the sense of having a similar biology, same culture, etc., is what forms the basis of this agreement and what it is we are disposed to judge.

The resulting picture is therefore not much different: if a particular case lies on the second-order equilibrium path of a basic constitutive practice, that case is constitutive of that practice as well, and hence of what is correct. That equilibrium point is determined by our stable disposition to judgement, and hence our stable dispositions to judgement indirectly determine, in each case, what our practice is. Here, the notion of stable disposition to judgement plays the role of 'naturalness' ('what we would all say') and each point on the equilibrium path the role of 'agreement'.

In the case of words like 'red', i.e. words that describe physical objects, a part of our practice are certain interactions with the external world—we learn the use of the term 'red' by being shown red things, looking at red things, fetching red things, speaking about red things and so on. The meaning of the term 'red' is constituted by our common dispositions to stable judgement, including judgement about the particular case, but *that* something is red depends on that very thing—that is to say, despite our agreement that this thing is *red* being constitutive of the meaning of 'red', propositions in which the term 'red' figures are nevertheless descriptions of empirical reality (at least in the paradigmatic case)—our agreement is merely constitutive of the meaning of the term 'red', not *that* any particular object is red. Simply put, given our training and natural make-up, we will not have the stable disposition to judge that a red ball is present, unless there is a red ball present. For more discussion on this point, again see Chapter 4.

Following Wittgenstein in explaining 'true' mathematical statements as being entirely defined in language is therefore easy enough: simple mathematical statements, e.g. of the form $57+68 = 125$, are true because they are correct without any contribution from the external world—e.g. those that lie on a second-order equilibrium of a basic constitutive practice of using a mathematical term. We learn a technique of calculating, for example addition or counting, by training and seeing examples, and extend them into novel cases by analogy. Those outcomes that lie on the second-order equilibrium of the basic constitutive practice of adding are those which are correct, and since mathematical statements are not descriptions of an external reality, but entirely defined in language by an appeal to the technique, the correct statements are the true mathematical statements.[2]

This picture does involve 'agreement about the particular case' in the sense detailed in Chapter 4. The idea is this: we have stable dispositions to judgement about very simple cases, like the sums of small finite numbers, and how to perform the carrying operation. A calculation is composed of a series of small such steps, no matter how complicated, and hence we can have disposition to judge about every possible step in any calculation, no matter how long. These dispositions to judgement combine through the game-theoretic framework as a second-order equilibrium path, and hence our agreement about each case, understood as a point on the second-order equilibrium path, is constitutive of the practice.

Logical inference can be given a similar treatment: a correct inference is one that lies on the second-order equilibrium of the basic constitutive practice of inferring: if we accept the statement $p$ and the statement that $p$ implies $q$, then given our training, the judgement that the statement $q$ is true will lie on the second-order equilibrium path of the practice of inferring, and hence correct. Since there is no appeal to anything outside of our practice, this account avoids Quine's regress problem for orthodox conventionalism: there is no need to look outside of the practice for an unreduced notion of consequence which is itself not conventional. It is convention all the way down.

The claim is of course *not* that if the variables stand for empirical statements,

2. Notice that it is not necessary that there exists a linguistic representation of this correctness in the form of propositions: we might have learnt the technique of adding and never utter propositions of the form $n + m = p$. See e.g. *RFM* I, App. §4.

then the conclusion will be *true* because of our consensus. That would be a case where the argument from worldly fact would in fact bite. It is rather that inferences of that *form* are correct because of the basic constitutive practice of inferring, and hence if the statements in question are not empirical statements, but, for example, uninterpreted variables or mathematical statements (which on this view *are* not descriptions of an external reality, but themselves true because of our consensus), then the practice is the only criterion of correctness. If the statements are empirical statements, however, there is no guarantee, on this account, that our basic constitutive practice of inferring will hit the mark—correctness in inference and truth could conceivably come apart. There is, on this view, no guarantee that our mathematical and logical concepts are empirically adequate.[3]

This does not mean, however, that our mathematical or logical concepts themselves have empirical content, as it does not follow that any particular empirical statement can *falsify* a logical or mathematical statement, or even our practice. Suppose for instance that we had in fact adopted Kripke's quus function for our arithmetical needs. When we then discover that the empirical proposition that we express in our counterfactual language as "57 oranges added to 68 apples are five pieces of fruit" is false, that has not falsified the *arithmetical* proposition "$57 \oplus 68 = 5$". It is still true that $57 \oplus 68 = 5$, even if the empirical proposition that we arrive at via the arithmetical one is false. This discovery could therefore at best suggest the adoption of a new practice of arithmetic, and we might, for instance keep using Kripke's quus function for other purposes, but adopt a new practice to

3. Cf. also *RFM* III, §33:

> It can be asked: how did we come to utter the sentence '$p \supset p$' as a true assertion? Well, it was not used in practical linguistic intercourse,—but still there was an inclination to utter it in particular circumstances (when for example one was doing logic) *with conviction*.
>
> But what about'$p \supset p$'? I see in it a degenerate proposition, which is on the side of truth.

In the terms of my account here, we can understand this remark as saying that we are in fact disposed to judge that sentences like '$p \to p$' are true, because they have much in common with empirical statements of the same form, but the circumstances in which we do so are radically different: in one case the only determinant of truth is our practice, and in the latter there is the world to content with.

For further discussion of equating correctness in mathematical practice with mathematical truth, see the next section.

count objects. Both practices might continue as they are, living side by side, each with their own internal correctness conditions.[4]

In the case of logic, we can, for instance, suppose that we have by a complicated process of deduction deduced in some formalisation of classical logic that $\varphi \rightarrow p$ where $\varphi$ is some complex formula. Further suppose that every step in the proof lies on the second-order equilibrium path of our practice of deduction, and hence that this statement is in fact correctly deduced, according to this account. If we were to interpret $p$ and all the constitutive atomic formulas of $\varphi$ as empirical statements where the empirical statements that form the premises of the deduction are all true, there is no guarantee that the empirical statement we chose to replace $p$ is also true. The world might not cooperate with our practice, but that does not show that our practice was *incorrect* by its own lights—even if we would, most likely, but not necessarily, revise it if this were to happen.

It follows from the account of mathematical statements given above, that if the propositions in question are mathematical statements, then our basic constitutive practice of inferring a mathematical statement from simpler ones *is* what constitutes the truth of a more complex mathematical statement, since there is no reality that might throw us off track—nothing to constitute the truth of the statement in question other than perhaps other parts of our mathematical practice.

What should we then say about more complex mathematical statements, e.g. those which are not simple calculations or similar? In the *Lectures*, as we saw in the previous chapter, Wittgenstein adopted the counter-intuitive position that in such cases, our agreement about the particular case, e.g. in the case of Goldbach's conjecture, is still what determines correctness, and hence mathematical truth, as well as claiming that believing that Goldbach's conjecture will be proven *is* to believe that we will find it most natural to extend our mathematics in that way. If we adopt the version of radical conventionalism about mathematical and logical truth just outlined, and how we have interpreted Wittgenstein's notions of agreement and naturalness, this picture quite naturally follows. Here then, the most natural way would be to say that a complex mathematical statement is true if and only if it is a correct inference from other mathematical statements—correct, in turn, because of the game-theoretic interaction between our stable dispositions to

4. See Matthíasson, forthcoming for further discussion on this point.

judgement, and hence true.

For example, consider the claim that there is an infinity of primes. A simple proof runs as follows:

**Theorem 7.1.1** (Euclid)**.** There exist infinitely many prime numbers.

> *Proof.* Consider a finite list of prime numbers, $p_1, p_2, \ldots, p_n$. Let $P$ be the product of all the numbers in the list, such that $P = p_1 \times p_2 \times \ldots \times p_n$. Let $q = P + 1$
>
> Either $q$ is prime or not. If $q$ is prime, there is at least one more prime number which is not on the list. If $q$ is not prime, however, then by the definition of prime number, there exists some prime number $p$ such that $p$ divides $q$. If $p$ were on our list, then it would divide $P$ as it is the product of all prime numbers on the list. Since the difference between $P$ and $q$ is 1, $p$ would have to divide 1, and hence not be a prime number. $p$ can therefore not be on the the list.
>
> In either case, there is exists a prime number which is not on the list, and since our list was an arbitrary finite list, there must be an infinite number of primes. □

Here, the proof proceeds by simple inferences from definitions, other logical truths or previously accepted statements (e.g. that if $n$ divides $p$ and $q$, then $n$ divides the difference of $p$ and $q$). If we agree that a correct inference is one that lies on the second-order equilibrium path of the basic constitutive practice of inferring and that the definitions and less complex statements are likewise true because they lie on such an equilibrium path, then the conclusion is true for the same reason. Our agreement about the particular case, namely that there are infinitely many prime numbers, is constitutive of the correctness of that statement, and hence its truth, but we are driven to it by our prior dispositions to judgement ('what we find natural to say') derived from our training and form of life.[5]

5. What should we say about cases where we genuinely prove a complex mathematical statement, for example that there exist no numbers of a particular form, and then a counter-example is found? It seems that this account of proof cannot rule out such a case. That is true, but that simply means that our mathematics would be inconsistent if that situation would arise—there is no guarantee against inconsistency in rival theories either, so this cannot be taken to be an objection against radical conventionalism.

That is to say, it is not the case that our agreement about this particular case is, or even could be, *sui generis*—even if it is constitutive of the correctness of that case. The process by which we reach the conclusion that there are infinitely many primes is essential in forming our stable judgement that there are, and hence into the determination of correctness. It would not be possible for us, given our training and actual mathematical practices, to directly form the stable disposition that there is a finite number of prime numbers, since our dispositions that drive us towards the conclusion of the proof would override it—the notion of *proof* and practice of *proving* is therefore constitutive of the correctness of the statement that there are infinitely many primes.

Perhaps the point could be better seen with a simpler example. Suppose a community of agents that has undergone the same kind of training in arithmetic as we have and otherwise shares our form of life has *prima facie* judged that $591+21 = 70$. This judgement would not be stable for these agents, because they would *also* have the stable dispositions to judge that $1 + 1 = 2$, $2 + 9 = 11$ and have a stable disposition to carry the one in the usual way, resulting in the judgement that $591 + 21 = 612$, via the technique of calculation. This judgement, on the other hand, would be stable and override the *prima facie* judgement.

Hence, there is a sense in which we *are* compelled to accept the conclusion of the proof of the infinity of primes, but it is not because otherwise we would not correctly describe the mathematical reality or something like that, but because we only have the stable dispositions to extend the proof in that one way, given our basic constitutive practice of inferring. That one way is fixed in advance, if our dispositions are, and hence, to switch examples, there is also a sense in which Goldbach's conjecture is simply true or false, before the proof is found—given our mathematical practice, we can only have one stable judgement about that case.

If, on the other hand, we were to have different dispositions to judgement, we would thereby be using different concepts (since the second-order equilibrium path would be different) and therefore, what we would then be disposed to judge would be correct relative to those practices. Logically speaking, there *could* therefore be such a one-off, *sui generis* judgement which would be constitutive of our concepts, and hence correctness (like Wittgenstein's '$12 \times 12 = 143$' example) but in our actual mathematical practice that is not the case, and hence such one-off

judgements do not play a role.

The claim is of course not that there are not intra-mathematical reasons for the truth of the claim that there are infinitely many primes. That is of course the case, and those reasons are what the proof expresses. The claim is rather that *what* we take as mathematical reasons is given by the mathematical practice. The conjecture that there are infinitely many primes is correct, because it was correctly inferred from simpler statements, and the correctness of those statements, as well as the inference, is given by the second-order equilibrium of our basic constitutive practices—if we had had the stable disposition to judge otherwise, we would, as it were, have had a different mathematics.

**Equating mathematical truth with correctness in mathematical practice** The key move in the preceding account of mathematical truth is identifying it with correctness in mathematical practice: a mathematical statement is true if and only if it lies on the second-order equilibrium path of a basic constitutive practice of using a mathematical term. This claim runs counter to a very influential argument in the philosophy of mathematics, often attributed to Frege (see e.g. Linnebo 2017) and hence referred to as the *Fregean argument*.[6] The argument runs as follows. Take some true mathematical statement involving an existence claim, e.g. the proposition that there are four prime numbers between 1 and 10. If it is true, which it is, there are four prime numbers that have the property of being larger than 1 and smaller than 10. It follows, one might easily say, that there are prime numbers, and since prime numbers are numbers, it likewise follows that there are numbers.

It seems that anyone who agrees to the truth-value of the statement that there are four prime numbers between 1 and 10, must therefore also be committed to the existence of numbers. These objects, the numbers, a proponent of this argument might continue, cannot be pointed to, nor seem to have come into existence at any point in history and do not have any material effects on anything. It is then natural to suppose that they are not concrete, as well as being atemporal and acausal, in other words: abstract (and for similar reasons, independent from human language

6. Among recent philosophers who have endorsed this argument are e.g. Resnik (1981; 1997), Maddy (1990), Shapiro (1997), McEvoy (2012) and Marcus (2015).

and practices).

The following is a more rigorous breakdown of the premises of this argument:

(1) The singular terms of mathematical statements refer to mathematical objects, the predicates to properties of such objects and the first-order quantifier range over the objects.

(2) Mathematical statements are capable of being true or false, and those accepted as theorems are true.

(3) If a statement is true and quantifies over a range of objects, it is ontologically committed to the kind of objects its quantifiers range over.

(C) There are mathematical objects.

I will not evaluate this argument fully here but I want to note a few things. First of all, each of the premises has a strong intuitive appeal and—perhaps after the filling out of some details—it is deductively valid. The first premise (1) is perhaps the most contentious of the bunch. It seems, however, that if we reject it, we lose a certain uniformity in our semantics. If the terms in mathematical statements do not refer to objects in the same way as our statements about ordinary objects do, it might seem that other important semantic properties change as well, in perhaps unacceptable ways. For instance, it seems intuitive to say that if a statement about concrete objects is true, then this is so in virtue of how things stand with these objects. The statement 'the teacup is on the table' is true because of how things stand with the teacup and the table, and the fact that the word 'teacup' refers to a particular teacup, 'table' to a particular table, etc. If we accept that, it might further seem that if that is not what mathematical truth is, it is not really truth at all, or as W. D. Hart puts it, we might think

> that no sentence of a developed body of sentences is true unless there are in the offing singular terms referring to objects; truth requires reference to objects. (Hart 1991, p. 90)

If we do not accept (1) it therefore seems that the truth of ordinary sentences about physical objects and the truth of mathematical statements is not the same kind of

truth. I will refer to this desire to keep our semantics uniform as the *uniformity requirement*.

In Wittgenstein's works, there is quite a lot of tension on this point. Most of the time, he seems to be fine with speaking of true mathematical statements, for instance in the passage in the *Lectures* where he claims that true mathematical statements are completely defined in language (*LFM* XXVI, p. 249. See last chapter for discussion). He also, however, often seems take the view that it is not *essential* for mathematical statements to be expressed as propositions nor their being truth-apt. It also often appears that Wittgenstein goes in for a deflationary or minimalist theory of truth whereby '$p$ is true' is simply the same as $p$.[7] Kripke describes Wittgenstein's view as follows:

> We call something a proposition, and hence true or false, when in our language we apply the calculus of truth functions to it. That is, it is just a primitive part of our language game, not susceptible of deeper explanation, that truth functions are applied to certain sentences. (Kripke 1982, p. 86)

Accordingly, mathematical statements *are* true or false, because we use them in that way, and not because the describe mathematical reality. On this view, reference to objects is just not necessary for truth, *contra* Hart.

There is not space here to develop a fully-fledged theory of truth, but I would suggest that there is a natural way for the radical conventionalist to allow for mathematical statements to be truth-apt, while respecting the intuitive force of the Fregean argument. The radical conventionalist might simply say: the concept *truth*—like any other concept—is defined by a second-order equilibrium of the practice of using the word 'true'. We learn how to use the words 'true', 'truth' and so on, by examples and training, and mathematical statements are among those that form our dispositions regarding the use of the word 'true'. Mathematical statements, whether they refer to objects or not, are therefore truth-apt by definition, because that we call them 'true' is, among other things, constitutive of the

7. For this reading, see for instance Kripke 1982, p. 86, Dummett 1978, p. xxxiv. For a different reading and discussion, see Vision 2005. For Vision, Wittgenstein should rather be seen as adopting a 'no theory' view of truth.

concept *truth* on the radical conventionalist account of meaning. The same would go for other related, but philosophically important, concepts like *reference* or even *existence*: we learn the meaning of the words 'reference' and 'existence' in part by seeing them used in true mathematical statements, and hence that these concepts apply to mathematical statements is constitutive of the meaning of those very concepts.

The radical conventionalist would then not hesitate to assent to the statement that mathematical objects exist, as long as the underlying account is kept in mind, namely that a mathematical object is only said to exist if it figures in the right kind of true mathematical statement, and that *that* mathematical statement is only true if it lies on the second-order equilibrium path of a basic constitutive practice and does not function as a description of an external reality. We might say, as Wittgenstein does in the *Lectures*, that it is only when the 'wrong picture' is associated with these kinds of locutions that we have said something false.[8] The Fregean argument can therefore be fully endorsed by the radical conventionalist, without subsequently forcing the adoption of the platonist picture of mathematical truth and an abandonment of radical conventionalism itself, if we just apply the conventionalist theory of meaning to the important philosophical terms that figure in the argument itself. Hence there would be no special problem about equating mathematical truth and mathematical correctness for the radical conventionalist. That they are the same follows quite easily from other commitments about meaning that the radical conventionalist already has.

An objection to this might go as follows: The account just offered doesn't distinguish between (1) the use of 'refers' in the object language, i.e. from with in the practice, and (2) the use of 'refers' in the metalanguage, used to describe the practice. We might be able to account for why agents from within the practice assent to statements such as 'numeral refers to numbers' by appealing to the radical conventionalist account of meaning, but I, in my anti-platonist role as a philosopher,

8. See Chapter 6 for discussion. We, as philosophers, might also want to distinguish between different senses of 'exist' in light of this account. We could for instance say that an object only really exists if it does so independently of our practice. The radical conventionalist would of course agree that mathematical objects do not exist in this sense. However, if we make such distinctions, then the Fregean argument would have to be modified accordingly, allowing the radical conventionalist to reject it.

say in the metalanguage that there are no such objects as numbers to refer to. The Fregean argument is a challenge to that claim, and not to what the agents say in their practice.

This challenge doesn't quite work, however. The radical conventionalist claims that we, the philosophers reasoning about our own practice, learn the meaning of the word 'reference' ourselves in part by learning how to use true mathematical statements of the right sort. Hence, the statement 'numerals refer to numbers' is literally true for the radical conventionalist, both in the object language and in the metalanguage, since *that* mathematical statements refer is partially constitutive of the concept *reference* in the metalanguage too. The disagreement is about the *nature* of reference, not that mathematical statements refer. From the radical conventionalist point of view, platonists conceive of reference in general on the model of reference to material objects, while the radical conventionalist has a more deflationary notion, where the circumstances under which mathematical terms refer (or are said to exist) are fully exhausted by our mathematical practice.

In a sense the radical conventionalist would thereby have given up the uniformity requirement, as reference to physical objects and reference to mathematical objects is not the same thing in that they do not have the same function and are true under different kind of circumstances. However, there is a uniformity of a different sort: they are both a part of the same basic constitutive practice of using the word 'reference', and hence both instances of the same concept. There are not, on this view, two concepts of reference, completely disjoint. The two kinds of sentence can therefore be given the same kind of semantic treatment, even if the metasemantics is different (e.g. the kind of referential semantics the Fregean argument presupposes).

**But is it conventionalism?** In Chapter 4, I stated that I was not prepared to say that basic constitutive practices are conventions as such. There are several reasons for this. The first is that on most definitions of convention (Marmor 2009; Vanderschraaf 2018; D. Lewis 1969), conventions are *arbitrary*: there has to be, according to these definitions, some other convention that could have served just as well for some purpose. This is of course a resonable and intuitive, and perhaps even truistic, requirement. Since I'm analysing meaning as *constitutive practices*,

however, it is unclear what this would mean for this account. If our basic constitutive practice of using a particular symbol was different, then its meaning would change, and it would thereby not be the same practice and therefore not give the same meaning.

This notion of arbitrariness is therefore not entirely clear in this case: we cannot arbitrarily choose a different basic constitutive practice without *thereby* being engaged in a different practice. This is not the case for conventions in general, since e.g. driving on the left side of the road instead of the right is still driving—and each serves the purpose of regulating traffic equally well.

Another reason is that it is common for philosophers to claim that conventions are arbitrarily chosen *rules* (Marmor 2009; Wikforss 2016). On the present account, meaning and (constitutive) rules are explained by reference to constitutive practices, not rules. There is a constitutive rule in play, but one given content by the practice. The practice is therefore not *selecting* different constitutive rules (nor regulative ones) and there is no sense in which meaning is an arbitrary chosen rule on this account—a different practice gives a different meaning, and a different associated constitutive rule, even if the *symbol* we use might be the same. This rule is not really what is giving the correctness conditions of the concept in the first place, despite being trivially read of the practice and the possible use of it by the agents themselves to explain their own practice (i.e. a participant in the practice might *explain* to another the basic constitutive practice of using the term 'red' by referring to a rule).

Despite those considerations, the label 'radical conventionalism' is nevertheless apt for the position I am defending here. First of all, if we look at common definitions of conventionalism, e.g. those discussed in the previous chapter, many of them emphasise that on a conventionalist view, mathematical statements are either true in virtue of meaning or in some sense truths of language. The radical conventionalist view, as it is outlined here, certainly fits that description: a meaning of a term referring to a concept, on the account given here, is given by the basic constitutive practice of using that term. By equating mathematical truth with correctness in such a practice, mathematical truths *are* true in virtue of meaning, and given how the account is intended to explain meaning, truths of language as well.

Other definitions, such as Dummett's, claim that conventionalism is the view

that mathematical truths are not imposed upon us by reality or that mathematical statements are in some sense or another 'up to us'. On that score, the version of radical conventionalism offered here also fits the bill.

Nevertheless, the idea that conventionalism depends on the arbitrary shouldn't be completely ruled out, and even if basic constitutive practices are not arbitrary in the sense that another basic constitutive practice could have done just as well for doing what the practice is a practice of, e.g. in the case of adding, to add, they are arbitrary in another sense, namely in that it is not *necessary* that we should have adopted the basic constitutive practices that we did. First of all, we could have had some other adding-like practice instead of the one that we actually have without thereby having gone wrong, for instance the practice of quadding.[9] That, granted, would not have been the practice of adding, but given other practices and aims specified independently of the practice of adding, would have been possible—we *could* have been quadders, instead of adders, for example, and if so, we would not have been doing something wrong or incorrect.

Secondly, given the examples and training we receive in the use of certain terms, the fact that one basic constitutive practice is adopted and not another is not a necessary one: it is in a certain sense an empirical fact that we have these concepts, and not some others (although it does not follow *that* $2 + 2 = 4$ is thereby an empirical fact). Basic constitutive practices are therefore arbitrary in the sense that the external world cannot *force* us to adopt any particular concepts, and we could have, if contingent matters were different, adopted others. "Necessity rides on the back of contingency", as Gerrard quotes W. W. Tait as saying (Gerrard 2018, p. 159).[10]

Yemima Ben-Menahem makes a similar point (Ben-Menahem 1998). She points out that the rules of chess are in a certain sense arbitrary, since we could have easily have adopted any other set of rules when coming up with a chess-like game. However, that game would not have been chess, which is defined by its rules (or

9. Many of Wittgenstein's examples in the *Remarks*—described by Dummett as "thin and unconvincing (Dummett 1959, p. 333)"—as well as in the *Lectures*, are intended to demonstrate that point.

10. I found the turn of phrase illuminating, but Gerrard does not say where Tait is meant to have said this and does not provide a reference. I was unable to locate it myself.

sufficiently many of them). Such conventions she calls 'constitutive conventions'.[11] She concludes:

> It appears that we must distinguish between different senses in which a convention can be arbitrary. In one sense, a convention is arbitrary when it cannot be justified, that is, when its choice is not constrained by other conventions or by external facts; in another, it is arbitrary if it can be changed without changing the nature of the activity or the meaning of the expression under consideration. A constitutive convention, we noted, can only be arbitrary in the former sense. (ibid., p. 105)

If we accept this distinction between different kinds of arbitrariness, it is clear that basic constitutive practice are of the first kind. They are not arbitrary in the sense that we could have done something else in order to be engaged in the same kind of activity, but the choice of such a practice is not constrained by any external factors, nor are they strictly speaking constrained by other conventions. That does not mean, however, that certain practices are not more natural or obvious, given certain empirical facts or other conventions that we have.

## 7.2 The arguments against radical conventionalism

In this section, I will respond to a number of arguments against radical conventionalism. I will start by examining a number of influential arguments due to Michael Dummett, then move on to more recent arguments against the view due to Severin Schroeder. Finally, I will respond to Putnam's so-called 'consistency'-objection.

### 7.2.1 Dummett's arguments against radical conventionalism

In Dummett's two papers there are a number of forceful counterarguments to radical conventionalism. In the literature, these arguments have been taken to be quite definitive against the view. Here, I will argue that Dummett's arguments do not

11. See also Marmor 2009, Chapter, 2 for discussion.

succeed in undermining radical conventionalism as outlined here and that radical conventionalism remains a viable view.

Dummett's first definition of radical conventionalism, in his first paper on the subject, on which he bases his subsequent criticisms bears repeating, and runs as follows:[12]

> [The] logical necessity of any statement is always the *direct* expression of a linguistic convention. That a given statement is necessary consists always in our having expressly decided to treat that very statement as unassailable; it cannot rest on our having adopted certain other conventions which are found to involve our treating it so. (Dummett 1959, p. 329)

There are two aspects to this definition. The first is that any true statement of logic and mathematics is always a direct *expression* of a convention, which I have taken to mean that the truth of a given statement is not due to it being a *consequence* of other, more basic statements (even though such statements might nevertheless be justified or be known through inference) and there is no difference between simple mathematical statements and more complex ones, from the point of view of the convention—the ground for their truth is the same. Their truth, as I put it, is directly determined by our mathematical practices. It does not *follow* from our practice that a given statement is true, it is determined by it.

The second is that we *decide* that the given statement is necessary. In Chapter 5, I argued that the second aspect is of lesser importance and that rather than understanding the language of decision as involving an 'act of decision', we should rather understand it as a general expression of conventionalism, that the truth or necessity of the given statement is "up to us".

**The freedom of machines** The first objection I want to consider concerns rule-following in general, in particular what it is to follow a proof in mathematics.

Dummett understands Wittgenstein's argument somewhat as follows: In order for us to follow a proof, we must recognise how to apply the rules of inference, even if we explicitly formulated them at the start and agreed to them, as the adoption

12. However, see Chapter 5 for more discussion.

of a rule is in itself not the same as recognising when to apply that rule and so our agreement with the axioms does not in itself constitute an agreement with each step of the proof. If we have do have a proof, we would indeed say that whoever doesn't agree with the theorem it proves didn't understand it or the rules of inference used in it, but it must not, therefore, be the case that there was anything she said or did in advance that could have ruled out such a misunderstanding, either of the proof or the rules of inference. So, we are free to choose, at each step, whether or not to accept it, and there is nothing in how we formulate the axioms nor in our minds that could have shown whether or not we should have accepted it or not, and therefore nothing that can force us to accept the proof. If we do, we say that theorem is now necessarily true and will not accept any experience as counting against it, and this act was a new decision, and not merely making what we had already decided, and was until now implicit, explicit (ibid., p. 330).

Dummett's first reaction to this argument that he attributes to Wittgenstein is to point out that while this seems a natural thought in cases where we have not formulated our rules of inference clearly enough, there will be no ambiguity if we simply give a precise enough formulation of the rule. Dummett thinks that Wittgenstein cannot accept this answer because he is hostile to mathematical logic as a bad influence on philosophers (ibid., p. 330)[13] and that this remark "as it stands" is so plainly silly that "it is difficult to get a clear view of the matter".

To illustrate his point, Dummett asks us to consider Wittgenstein's example of adding. We train someone to follow an order of the form "add *n*" where we have only shown them a finite set of examples containing relatively small numbers and then give them a particular order, e.g. "add 1" and they add one for every number up to 99, two for every number from 100 to 199, three for every number from 200 to 299 and so on. For Wittgenstein, there is nothing, Dummett says, no fact about them (or their mind) or how they were trained that could have ruled out this aberrant interpretation of the rule (unless the training examples had extended further than the wrong application, of course, but then there would also have been fresh possibilities for misunderstanding). For Dummett, this is all well and good, if we take it to tell us something about intention and its nature, but as illuminating

13. Dummett's refers the reader to RFM IV, §48 for this remark but in the third edition it is at RFM V, §48.

something about mathematical proof, it seems to miss the target. For, can we not give precise formulations of rules, rather than teaching them by example in such a roundabout way and even program computers to consistently and mechanically follow rules, for example, increase a number by one or add two? Does that not show that we can follow the rules of inference in a unique way when reading proofs? Or as Dummett puts it, "A machine can follow this rule; whence does a human being gain a freedom of choice in this matter which the machine does not possess?" (Dummett 1959, p. 331).

This point is well taken, but it is hard to see why Dummett takes Wittgenstein to rule out his response on the basis of his animosity towards mathematical logic. Is is simply odd to take his animosity towards logic as a reason to reject Dummett's point about clearly proscribed rules and machines. First of all, Wittgenstein's remark does not appear in any discussion of rule-following or related matters (as far as I can see) and even if that were the case, claiming that logic has distorted the thinking of philosophers is no answer when it is pointed out that machines can follow rules. It would be, as Dummett himself puts it, a very silly reply and there is no reason to think that Wittgenstein's ever attempted (or would have) to make it when faced with an argument like Dummett's.

As for the two aspects of Dummett's objection, namely that Wittgenstein's argument does not apply if the rules if they are rigorous enough and that machines, that is to say computers, can follow a rule such as our rule for developing the series of natural numbers or adding two, should be easy to answer. The answer to the first is of course, as Dummett must realise, that the rule-following argument has nothing to do with how precisely the rule is formulated, and Dummett's only reply to that point, which he grants, is to say that this is a general point about meaning, but does not have any relevance to the philosophy of mathematics (ibid., p. 331). If the topic in the philosophy of mathematics under discussion concerns inference rules, that claim is simply unconvincing.

The reply to the second point is also quite straight-forward: machines do not follow rules in the strict sense of the term, but merely act in accordance with them. In my view, Kripke's arguments against Dummett are definite on this point (Kripke 1982, pp. 33–36). Kripke points out that the term 'machine' is ambiguous in various ways. We can first suppose that it refers to a computer program, a se-

quence of instructions to be executed by a computer. Here, the sceptic can simply say that the same problem arises as for the symbol '+': the program itself should really be interpreted to implement the quadding function, and not the adding function. If we want to say that the program considered as an abstract object should count, the sceptic can ask to which abstract program does the program I wrote down on paper refer to (to which program I intended to refer) and hence bring back all the other problems about rule-following discussed in Chapter 2.

Finally, we might consider that the machine is a physical machine, made of metal and gears (or perhaps silicon and transistors). In this case there are two problems the sceptic can exploit. First, the machine, as Kripke points out, is a finite object and can only compute finitely many numbers—for any actual calculating machine, including the one I'm typing this on, there are numbers that are simply too large for it to calculate. There are therefore infinitely many ways to extend the actual behaviour of the machine, and we cannot simply say that whatever it *would* do is correct, since that would rely on a question-begging *ceteris paribus*-clause. If we'd want to refer to the intentions of the programmer in this case, the reference to the machine would be superfluous. Second, we would lack a criterion for correctness if the machine malfunctions—a variation of the problem of error. Real machines break, and give the wrong answer on occasion, and so we either take whatever the machine does as correct, and there is no such thing as the machine malfunctioning, or there is some other criterion of correctness than the operations of the machine.

The rule-following paradox, and radical conventionalism about mathematics as a response to it, is really about how correctness in a mathematical practice gets established. Dummett's questions ignore that point.

**Dummett's dilemma** The next argument I wish to consider is a dilemma that Dummett presents for the radical conventionalist and one of the last arguments, he presents against radical conventionalism, but in my view one of the most compelling (Dummett 1993).

It goes as follows: If we think that nothing external to our practice nor what we have done in the past ("the concepts themselves" as "they have previously been specified") determines what we should do in the future, it seems that we have to

accept a very counter-intuitive picture of what a proof is. Namely, that on the assumption that a proof of some theorem is always composed of smaller little steps from the premises to the conclusion, it would then seem that if each statement is a direct expression of a convention, we either have to say that it is not determinate what counts as right when we take each of the little steps or that the combination of them is not transitive, that we can accept all the little steps and reject the whole proof. Dummett writes:

> Suppose the calculation in question is an ordinary addition. One of the rules that make up the computation procedure is that, if one of the two final digits is 7 and the other 8, you write 5 in the digits column of the sum and carry 1 to the tens column. To maintain that there is no determinately correct result of the calculation, you must say one of two incredible things. Either you must say that, until someone has done it, it is not determinate what would count as writing down 5 and carrying 1; or you must say that, although it is determinate what the outcome of each application of one of the constituent rules would be, it is not determinate what would be the outcome of a large but finite number of such applications. I do not know how many of the followers of Wittgenstein really believe either of these things; for myself, I cannot, and conclude that the celebrated 'rule-following considerations' embody a huge mistake. (Dummett 1993, p. 460)

In fact, I think that the radical conventionalist, while accepting the dilemma in a certain sense, can also reject it in another, more important sense. These two different senses depend on how we understand the notions 'determine', 'acknowledge' and so on. Dummett points this out indirectly, but he notes that the dilemma depends on an extra premise:

> He [i.e. Wittgenstein] was certainly right to observe that, for the most fundamental of the rules that we follow, there is nothing *by which* we judge something to be a correct application of them. It certainly does not follow from this that, if we never do make such a judgement in some particular instance, there is no specific thing that would have been a correct application: to draw that inference, you need a general

> internalist premiss, that there is nothing to truth beyond our acknowledgement of truth.(ibid., p. 460)

In effect, Dummett accepts that the rule-following considerations show that there is nothing (and I read it as: no fact, in the sense of factualism about meaning)[14] that can determine the correct application of a rule in advance but at the same time that this does not show that there is nothing which counts as correct or incorrect in applying them. And this is quite right on the radical conventionalist account offered here: there is nothing (i.e. no fact in that sense) that determines in advance what is correct, but the second-order equilibrium path of the practice does determine every case in advance, as soon as the dispositions are fixed. That is to say, the radical conventionalist does not accept Dummett's extra premise, and thus does not make the inference he claims is needed for the repugnant conclusion he draws from it, namely that there is nothing specific which is correct.

The radical conventionalist, however, does not place the source of the normativity outside of the practice itself, and so can still say that each arithmetical truth is a direct expression of a convention in the sense specified here (with the usual caveat that basic constitutive practices are not conventions as such). The radical conventionalist can then even accept Dummett's extra premise in a certain sense, namely with the caveat that 'acknowledge' does not mean here anything more than 'being conventionally true'—i.e. that there is nothing more to arithmetical truth than what the basic constitutive practice itself determines. Hence, the radical conventionalist accepts the dilemma in the sense that it really isn't determined before each of the little steps is taken what counts as right, in the sense that some fact or the concepts themselves as previously specified determine correctness independently of the practice, but at the same time maintain that there is something right or wrong in play.

In other words, if we understand 'determine' in a strong way, where there is some kind of rule, definition or meaning independent of human practices which settles *a priori* what the outcome should be, then the radical conventionalist accepts the dilemma and opts for the first horn: each step is undetermined in this sense. If, however, we take it in the weaker sense, that the practice itself—seen as the

14. See Chapter 2 for discussion.

complex, structured interplay between agents—can provide a criterion of right and wrong, there is 'determinately correct answer' in each case: the one where the basic constitutive practice is in equilibrium.

What can we then say about Dummett's example of addition? Suppose we're adding the numbers 8 and 7. What we are supposed to do, in carrying out this calculation according to one particular technique of adding is to write down 5 for the units and 'carry' 1 over to the ten's column. For Dummett, it is a consequence of Wittgenstein's view that it is not determinate that I should write down 5 and carry 1, but that's not what we should say. We should say: for every step in carrying out the calculation, the basic constitutive practice of adding determines what counts as a correct step. Hence, we cannot in practice accept each little step and reject the final outcome. We do not *call* anything but a particular thing an instance of 'carrying' and we have the standing to make this judgement because of the structure of our practice. We could, however, have had a different basic constitutive practice where each step was correct, but not the final outcome, but that would then be a difference practice with a different second-order equilibrium. Hence, given a particular constitutive practice, the scenario that Dummett envisions cannot occur.

In effect, the radical conventionalist *does* say that each truth is directly determined by the practice, i.e. by lying on the second-order equilibrium path of a basic constitutive practice. This is not in contradiction with there being a correct or incorrect answer at each step, since the correctness conditions are thereby constitutively determined.

**Mathematical surprise and the phenomenology of proof** The third of Dummett's objections I want to consider is related to the last one, and has to do with the phenomenology of proof. One of the biggest problems facing radical conventionalism as it is described by Dummett is that it simply does not seem to square with our phenomenology of doing mathematics and very natural intuitions we have when we are doing mathematics. It does not *seem* to us as if we're deciding what the outcome of either calculations or proofs are and we certainly don't tend to think that at every step of a proof were are merely making a new decision or stipulation of what now to accept as true, but rather that each step really follows necessarily from the previous ones, completely independently of what we

may think of it or what we might decide, that if we have understood the axioms and the rules of inference correctly, then we have no other intellectual choice in the matter of whether or not to accept the result or not.

We feel as if we are, as Dummett puts it, "mere passive spectators" to the proof and have no further active part to play when it has all been laid down. Because of this, we can always be surprised by the outcome of a long chain of reasoning in a way that talk of decisions seem to exclude. There could not, Dummett seems to be saying, be anything like mathematical discovery on this account, because each truth is up to us.

So far, I've repeatedly argued that the relevant notion of 'decision' is best seen metaphorically: when we take each little step in a mathematical proof we are following an analogy of what we have done previously, given our training etc. and that the 'decision' of which analogy is correct is 'taken' by appeal to the structure of the relevant basic constitutive practice. What we have done previously does not determine which analogy is correct, since we can always find different analogies which are consistent with what we have previously done, but we merely all do follow the same analogy as a matter of fact and the equilibrium of the convention provides the correctness conditions because that is constitutive of what is correct. The basic constitutive practice can determine the correctness conditions of a new case, without having been defined by a rule beforehand.

Let's look at a very simple toy example of mathematical surprise and see how this account would handle it. Suppose someone sets us the following problem: Imagine we have a rope that circumfers the globe at the equator, all 40 075 kilometres, and you now want to lift the rope so that it is one meter above the surface of the Earth while still circumfering the globe in the same place. How much rope do you need to add to your old rope to be able to make it?

Now, the most intuitive answer is that you need a lot, somehow proportional to the size of the Earth, but suppose I reason as follows. I know that the circumference of a circle is defined by $S = 2\pi r$ where $r$ is the radius. To calculate the

circumference after increasing the radius by one, I get

$$\begin{aligned} S' &= 2\pi(r+1) \\ &= 2\pi r + 2\pi \\ &= S + 2\pi \end{aligned}$$

I can thus conclude that no matter how large a circle is (and so in particular if the circumference is 40 075 kilometres) I need roughly 6.3 metres of extra rope to lift it up one meter.[15]

The result is certainly surprising (or it was for me when I first encountered the proof)—that no matter if a circle has the radius of a football or is the size of the Earth, increasing the radius by 1 always increases the circumference by $2\pi$; the rope could have been circumfering the visible universe and yet only 6.3 or so metres of rope would be needed if it were to expand by one metre. Of course, after seeing the proof, the point is obvious.

On Dummett's reading of Wittgenstein, what is meant by that 'decision' is that I could have freely chosen to adopt, for instance $\frac{1}{2}\pi$ as the result instead of the answer I presented here. But if we look at the example, what I actually did was to take a number of little steps, each of which is obvious (or rather: non-surprising), including the second to last step before the conclusion, which added up to a surprising conclusion.

What should the radical conventionalist say of each of the little steps? Remember, on this view, we learn a certain technique and go on as we do in new cases because we see a certain analogy with previous cases. This analogy is not logically determined as being the correct one, in the sense that all other analogies are ruled out *a priori* by an appeal to something that exists outside of the basic constitutive practice. At each step, however, there is both only one course of action that we, seeing the analogy of how to continue, would take and only one course of action that correctly finds the second-order equilibrium path of the basic constitutive practice,

15. Here it is of course necessary to distinguish between the mathematical proposition that $S' = S + 2\pi$ and the empirical question that a rope of a particular length will circumfer a particular circle. It is the business of our mathematical basic constitutive practices to determine the former, not the latter, as the argument from worldly fact would show.

Wittgenstein often speaks of mathematical statements as being rules we use to judge experience, and this example would be a good one to demonstrate that. See fotnote on p. 189.

but that, again, depends on the prior contingent fact that we all do see the same analogies. This fact, however, is a brute, empirical fact, and does not determine the outcome in the strong sense discussed above.

The radical conventionalist position thus *mimics* what it would be like to grasp the meaning of a mathematical concept and the rules of inference and follow them up as if they were independent of our linguistic practices but what is actually going on is that the agent is acting on their own dispositions to judgement, where the correctness of each step is fully explained by an appeal to our linguistic practices—without falling into the orthodox conventionalist trap of relying on logical notions given independently of the practice. The practice, and its role, is opaque to the agent taking part in it. We, as individual rule-followers, are thus 'passive spectators' in some sense, while the the whole linguistic community, through the basic constitutive practice which 'decides' the right way to proceed (with 'decision' and 'acceptance' understood metaphorically)—in the sense of picking out one second-order equilibrium path of the relevant basic constitutive practice out as correct. In this way, radical conventionalism *can* account for both the phenomenology of proof *and* the element of surprise in mathematics.

We can put the point slightly differently: when I'm considering what the concept requires of me, for instance when reasoning from the equation $S = 2\pi r$ to the equation $S' = 2\pi(r+1)$ if one is added to the radius, the boundaries of the concept are determined by the basic constitutive practice itself in each case (i.e. why this is allowed), but when I consider it in isolation, it seems that what I'm actually doing is reflecting on the concept itself, given independently of the practice, when what I am doing is taking part in a basic constitutive practice. This conventionalism is thus still radical, since there is no way to give a specification of the relevant concepts in advance or independently of the linguistic practice, even though in each case the practice determines only one path (because only one analogy of how to continue is in equilibrium) and can thus lead us to a surprising result. If I had gone on differently at each step I would have been reasoning incorrectly because I would not have done what the basic constitutive practice, so structured, required of me—I would have missed the equilibrium—and thus not taken part in that very practice. It would nevertheless *seem* to me that the practice plays little to no role.

Talk of 'decision' should thus be understood as meaning that there is no need

to refer to anything outside of the basic constitutive practice to account for the correctness of each step, not that each person or community is somehow free to decide whatever they want as the correct answer. The community could have adopted different basic constitutive practices, and hence had different concepts, but as long as the dispositions of the agents are fixed, we, both thought of individually and as a whole, are constrained by the practice itself, but that does not mean that the view is not conventionalist throughout: each truth is nevertheless a direct expression of the convention.

There are therefore a few reasons why we could not have accepted e.g. $\frac{1}{2}\pi$ as the answer to our puzzle, that is to say, to decide that this would be the correct answer in Dummett's sense. First of all, considered from the point of view of an individual participant in the practice, we would have been in contravention of our basic constitutive practice, which recall, does not depend on any individual decisions to begin with, since the every step lies on such an equilibrium path and the last step to the result as well. The result $\frac{1}{2}\pi$ would simply have been incorrect, given the practice.

Furthermore, we would not be able, from a psychological point of view, to just accept any piece of reasoning leading to this conclusion as the right analogy of how to continue, since that is not how we are disposed to act given our training. And hence, we would be in the (psychologically) impossible situation of either accepting the conclusion and not accepting the reasoning or not accepting the conclusion and accepting the reasoning, if we accepted $\frac{1}{2}\pi$ as the right answer. There would have been *some* analogy of how to continue that would have fit with what came before and that $\frac{1}{2}\pi$ was the right result, but it would not have been what *we* call the correct analogy—and that is determined by the basic constitutive practice and our stable dispositions to judgement which act as inputs to it. We are simply not disposed to conclude from the reasoning just displayed that $\frac{1}{2}\pi$ is right, and hence our basic constitutive practices ultimately preclude that answer.

Finally, our mathematical practices are mediated through *techniques*, such as the technique of adding or counting, and we do not necessarily have access to what our stable disposition to judgement is in any given case of sufficient complexity—in order for that to be manifested it is necessary to actually perform the calculation in question, and follow up on all the little steps. That process leads us to certain judge-

ments which get their correctness conditions from the basic constitutive practice, and that is what makes mathematical surprise possible. Our initial disposition, to say that some enormous amount of rope would be required, does not count as inputs to the basic constitutive practice, as that is not stable. Only our final disposition to judgement, manifested *after* the proof has been carried out, is (and it is so because of our training).

For an extended discussion of a similar point, see below.

**The case of elementary calculation** The next argument against radical conventionalism I want to consider concerns independent criteria for the same statement, e.g. in elementary calculations. We can imagine, Dummett says, that there were people that counted like us, but do not have the concept of addition. If one of these people, say $T$, had counted five boys and seven girls in a classroom, she would find out how many children were there in total by counting all of them. She might therefore be prepared to say on some occasion that there were five boys, seven girls and twelve children, and later that there were five boys, seven girls and thirteen children. In this case, her criteria for having miscounted would be noticing having counted a child twice and the like.

Now suppose we teach $T$ how to add and she subsequently comes to believe that whenever there are five boys and seven girls present, there are twelve children present. It would be quite right, Dummett thinks, to say that under these circumstances $T$ has adopted a new criterion for saying that there are twelve children in the classroom and new criterion for saying that she has miscounted (e.g. when the addition and the counting don't match up). Now, $T$ does not need to have noticed any mistake in order to come to believe that she miscounted—counting five boys, seven girls and thirteen children is enough.

But, Dummett says, we should say that even before $T$ learnt the principles of addition, it would have been the case that if she had counted thirteen children, she would have been wrong according to criteria that she herself already had. She must have, Dummett says, made a mistake in counting, and if she did make a mistake, there was something which she did, which is such that had she noticed it, she would have admitted that she did in fact miscount. Hence, the criterion of having miscounted introduced by addition isn't completely independent of what came be-

fore. The point is, I believe, that on Dummett's reading of radical conventionalism, *T*'s learning to add should have introduced criteria for having miscounted in that given case, completely independent of the criteria for having counted correctly, a reasonable assumption, given the definition of the view, but an implausible conclusion, and hence the case is a *reductio ad absurdum* of radical conventionalism.

However, I do not think that this argument succeeds. On the view as it is presented here, it will be the case that given *T*'s training and the basic constitutive practice of counting that she belongs to, that the correctness conditions of each step in counting the objects will be fixed, even if the number of them is extremely large (see Chapter 4 for discussion on that point). If *T* takes part in that practice, then there will be some step she got wrong if she got the result that there were thirteen children, and if she is like the other agents in the practice, there would be something that she did wrong, and hence possible for her to notice as having made a mistake. That just follows from the role training plays in the determination of meaning in a basic constitutive practice.

What about the point that the criteria would have to be completely independent according to the radical conventionalist? I accept Dummett's point that this would be highly implausible, but fortunately, I do not think it follows. First of all, the basic constitutive practice of adding *does* depend on the basic constitutive practice of counting in a meaningful sense: we cannot learn how to add without learning how to count, and our techniques for adding necessarily involve the basic constitutive practice of counting. It is therefore not the case that on the radical conventionalist view, these two criteria are completely independent. However, from the point of meaning-determination, they are: both statements, the one involving counting, and the one involving adding, are correct because they lie on their respective second-order equilibrium paths, and these are independent from each other.

The two practice depend on each other, because one of the ways we get our stable dispositions to judge that we have added correctly is by counting, and hence the way that the second-order equilibrium path for adding is determined depends on the second-order equilibrium path for counting, but they are independent in the sense that both equilibrium paths are determined on their own. In Chapter 4, I used Kripke's example of an algorithm for addition to demonstrate this point: we

might have a fixed practice of counting, but any algorithm for addition dependent on counting might be interpreted deviantly while holding our practice of counting fixed. We still need the basic constitutive practice of adding to play a role, even if we need other practices to be able to learn *that* practice.

**Communication breakdown** Before leaving Dummett's objections behind, I want to consider a brief one. Dummett points out that if we accept that we can simply make a decision at each point of which statements to accept as necessary—without regard for the statements we have already accepted as such—communication, and thus the use and point of language, would be in danger of breaking down completely. Dummett writes:

> Wittgenstein's quite different idea, that one has the right simply to lay down that the assertion of a statement of a given form is to regarded as always justified, without regard to the use that has already been given to the words contained in the statement, seems to me mistaken. (Dummett 1959, p. 337)

The reason, Dummett says, is that any mathematical statement seems to affect any other, and thus it is impossible to give an account of the meaning of words without giving an account of the whole language, and thus on Wittgensteins's account we will be unable to account for the meaning of *new* statements we have not encountered before.

The premise Dummett seems to accept here, which might explain his talk of decision more generally, is that *each* individual rule-follower is free to choose how to follow a rule. That premise, on the account given here, is of course false: there is no act of decision involved, particularly not by individual rule-followers. Hence, there is no particular problem about communication on this account, nor about new statements. There is no reason to suppose that we cannot explain the productivity of natural language by compositional semantics, for instance, with the caveat that the meaning of new statements wouldn't be grounded in such semantics, but rather that agents could *know* the meaning of each statement by such means and that their training, and hence how the concepts are constituted, relies on such compositionality.

Compositionality would then be analogous to techniques in mathematics; it is a way for agents taking part in the practice to discover what their stable dispositions to judgement are about a particular case, and so an explanation of how they can know what propositions they have never heard before mean, but also, because compositionality plays a role in our training and how we learn words in the first place, our stable dispositions to judgement are shaped by compositionality and thus compositionality plays a constitutive role in meaning determination, without being itself what determines meaning.

### 7.2.2 Schroeder's arguments against radical conventionalism

A more recent critic of radical conventionalism is Severin Schroeder. In a recent paper (Schroeder 2017a), he has advanced a number of arguments against radical conventionalism, in particular against its coherence. Here, I respond to those arguments.

**Failure of community solutions** The first of Schroeder's arguments against radical conventionalism I wish to mention is more an argument against community solutions in general than radical conventionalism in particular, but since radical conventionalism is a community view, it bears mention. The objection itself is fairly short, and goes as follows:

> As a response to Wittgenstein's rule-following problem the community view is a complete failure. For if it cannot be fixed in advance what in a given case is a correct application of the concept '+2', then it is equally impossible to fix in advance what in a given case is to count as 'community agreement' (RFM 392c). Both are on exactly the same footing as instances of Wittgenstein's problem: How can a *general* concept determine its *particular* applications? (ibid., p. 94)

Schroeder is, I believe, absolutely right on this score: these problems are the one and the same.

However, this is only shows that these problems must be solved simultaneously, and that, I claim, is what the solution developed in Chapter 4 does. On that solution, when the agents' stable dispositions to judgement is fixed, it *is* fixed in

advance what is to count as community agreement, namely whatever lies on the second-order equilibrium path of the basic constitutive practice in question. By identifying correctness with community agreement, the solution fixes in advance what counts as correct *by* fixing what counts as community agreement in advance. It is by identifying concepts with second-order equilibrium paths that a general concept *can* determine its particular applications.

Schroeder's reference to the *Remarks* is worth discussing as well, however, since it goes against my interpretation of Wittgenstein. It goes as follows:

> A language-game, in which someone calculates according to a rule and places the blocks of a building according to the results of the calculation. He has learnt to operate with written signs according to rules.—Once you have described the procedure of this teaching and learning, you have said everything that can be said about acting correctly according to a rule. We can go no further. It is no use, for example, to go back to the concept of agreement, because it is no more certain that one proceeding is in agreement with another, than that it has happened in accordance with a rule. Admittedly going according to a rule is also founded on an agreement. (*RFM* VII, §26)

This remark isn't entirely clear, but Wittgenstein seems to be saying here that while correctness in rule-following *is* founded on agreement, there is not use in investigating any further what that amounts to.

I do not doubt that this was Wittgenstein's considered view. However, as I argued in Chapter 1, it is unclear how this kind of quietism solves the problem of rule-following and what the emphasis on agreement then amounts to—why appeal to agreement at all, if there is no use in going back to that very concept in providing an explanation of rule-following? Wittgenstein has therefore given up on solving his own problem, rather than providing a solution, and there is no reason for us to take it on his authority, in the absence of further arguments to that affect, that it cannot be solved.

Does that mean that we should abandon the interpretation of Wittgenstein's philosophy of mathematics outlined in the last chapter? I don't think that we necessarily should, depending on our goal: it is doubtful—for better or worse—

that Wittgenstein's remarks on mathematics contain only one coherent philosophy of mathematics (and perhaps it contains none). Nor would we expect that to be the case, as neither the *Lectures* nor the *Remarks* were prepared for publication by Wittgenstein, or even vetted by him.

Furthermore, my reading of the *Lectures*, outlined in the last chapter, does not attribute to Wittgenstein any particular analysis of what agreement amounts to, unlike the account which is offered here, and hence the reading offered in the last chapter does not directly contradict the passage just cited, despite the heavy emphasis on agreement being constitutive of correctness. The account offered here goes further in that respect.

For those reasons, I prefer to state my case as being inspired by Wittgenstein, or being Wittgensteinian, rather than being Wittgenstein's as such.

**The circularity argument** The second of Schroeder's arguments against radical conventionalism I wish to discuss is in my view one of the most serious objections to the view, one that is implicit in much of Wittgenstein's writings on mathematics and briefly alluded to in the last chapter. We can call this the 'circularity argument'.

Schroeder's version of the objection runs like this: the mark of conventionality, as he puts it, is that the "standard of correctness is constituted by social agreement" (Schroeder 2017a, p. 94). Hence, Schroeder says, criticisms of deviation only need to reference that social agreement. And indeed, this is so for cases such as which number comes after 6 in the order of the natural numbers, or maybe simple cases like the proposition that $2 + 2 = 4$. We can understand what it is for us to have conventionally agreed that 7 comes after 6, and how we would justify that this is correct: by pointing to that agreement. But this is not the case in general. We have not agreed, for instance, that $135\,664 + 37\,863 = 173\,527$. And the reasons we give for this being correct is not that everybody accepts it, but rather the procedure of calculating. We have never encountered this sum before (presumably), nor does it figure in our teaching of sums, nor written in any arithmetic books and so on. In fact, it is just an empirical fact that there is no agreement about sums we have never before encountered and my accepting a given sum now has no bearing on whether future mathematicians will ever accept it, and they will certainly not

use my calculations as a justification for the correctness of theirs. There is only agreement, Schroeder says, on the general rules of addition, not particular sums.

Schroeder does not present his argument as a circularity argument, but rather as an empirical fact about mathematical practice, but I think it is useful to look at it from that angle as well. In *RFM* VI, §16, Wittgenstein writes:

> If it is not supposed to be an empirical proposition that the rule leads from 4 to 5, then *this*, the result, must be taken as the criterion for one's having gone by the rule.
>
> Thus the truth of the proposition that 4 + 1 makes 5 is, so to speak, *overdetermined*. Overdetermined by this, that the result of the operation is defined to be the criterion that this operation has been carried out.
>
> The proposition rests on one too many feet to be an empirical proposition. It will be used as a determination of the concept 'applying the operation +1 to 4'. For we now have a new way of judging whether someone has followed the rule. (*RFM* VI, §16)

Wittgenstein seems to be saying that the result of the calculation $4 + 1 = 5$ is the criterion that the rules of arithmetic are being followed, that the criterion for having correctly added the numbers 4 and 1 is that the result is 5. That's all well and good for a calculation like that, but for higher numbers, we run into a circle. We do not know if $345 \times 112 = 38\,640$ or not, and the method we use to verify this is to calculate. If Wittgenstein is right, the calculation is a way to discover the result *and* the criterion for the calculation having been correctly performed.

That might sound odd in itself, but matters are even worse if we suppose that our agreement is what determines correctness. If our agreement is meant to be the criterion for the correctness of the result 38 640, that agreement would have to exist in some sense before we perform the calculation, otherwise that result could not be the criterion for the calculation having been performed correctly. But we need to perform the calculation to reach that agreement in the first place.

In other words: If, we might say, mathematical correctness is just a question of nothing else than $345 \times 112 = 38\,640$ being agreed upon as being correct, how

would we then ever be able to declare some outcome either correct or incorrect, if we have no way to move from what we've said before to new and new cases? I can't tell if $345 \times 112 = 38\,640$ is correct or not and can therefore not tell if I should call that the correct outcome of the calculation or not. The calculation, on the other hand, is the *criterion* I use for determine that this is the right outcome, and can thus not without circularity be used as a criterion for the correctness of the calculation that same calculation. This is circular, because we need to perform a calculation to learn that $345 \times 112 = 38\,640$ and yet our agreement is supposed to be the criterion for the calculation having been correctly followed. How can that possibly be?

Fortunately, we can give a reply to both versions of this objection. First of all, to address Schroeder's version first, it is not the case that our agreement is constitutive of correctness, if we understand 'agreement' as opinion or anything of the sort. It is not the case that $345 \times 112 = 386\,40$ is correct because it is the result of some kind of vote where the majority carries the day. It is not agreement in that sense.[16] Rather, we are analysing 'agreement' game-theoretically as a point where our stable dispositions to judgement are in equilibrium, and this kind of agreement does not need to have been explicitly manifested or articulated to be real, nor to we point to it as justification when we act. The agreement is an agreement in dispositions to judge, not in saying aloud about a specific case: "Yes, I agree" or anything like that.[17]

Secondly, to address both versions, it is absolutely true that we find out that

16. Another author who gives a similar objection to radical conventionalist interpretations of Wittgenstein's philosophy of mathematics is Steven Gerrard (2018). For him, the radical conventionalist is tempted to "misinterpret " Wittgenstein's notion of agreement

> as meaning that the truth or falsehood of mathematical propositions is a result of some sort of mathematicians' annual convention: they vote on whether 81 + 81 equals 162 or 163, and the winning proposition goes into the archives. (ibid., p. 171)

According to Gerrard, the agreement in question isn't agreement about the particular case, but one that provides a framework without which "there could be no such thing as correctness or incorrectness" (ibid., p. 172). What I hope to have shown in this thesis is that radical conventionalism does not presuppose this misinterpretation, and is not only compatible with the 'framework' view Gerrard (in my view, correctly) describes, but rather fills out its details.

17. See Chapter 4 and Chapter 6 for discussion, as well as *RFM* VI, §49: "The agreement of humans that is a presupposition of logic is not an agreement in *opinions*, much less in opinions on questions of logic".

$345 \times 112 = 38\,640$ by the process of calculation and that there is social agreement about the method of calculation. But that was precisely the problem that the account is meant to solve: how can it be that to follow a rule is a practice, based on agreement? After all, when we learn a method of calculation, the teaching underdetermines which practice it is, and if our agreement isn't about the specific case, then what? (If we agree that *this* way is correct, what is this *this* that we agree on?)

On the account offered here, there is social agreement about *every* step in performing the calculation and that is constitutive of the concept we are employing, and hence the correctness of those steps. The calculation is composed of a number of small little steps, each of which is correct because they lie on the second-order equilibrium of a basic constitutive practice of adding—and that is the reason why we all stably judge *that* $345 \times 112 = 38\,640$, and hence why we settle on that particular outcome. We do not have an immediate judgement that $345 \times 112 = 38\,640$, but have other, stable dispositions that through the technique combine into that judgement.

It is not a contradiction on this account to say both that we find out *that* $345 \times 112 = 38\,640$ and at the same time that our agreement is constitutive of the correctness of that sum. It is rather what we would expect. How does that answer the circularity objection then? It is by having a stable disposition to judge about each step in the calculation that the correctness of the outcome is established, and those stable dispositions are shaped by how we learn the relevant concept, and in the case of a large sum, that is by performing a calculation. These dispositions, because of the shape of the account, are fixed as soon as we've acquired the concept, but that doesn't mean that we automatically know what we are stably disposed to say in advance, even if we have those dispositions. We might, however, have all kinds of prior dispositions regarding that sum, but because of our training, only our dispositions mediated through the calculation will be stable. We will only agree that one way of proceeding through the calculation is correct.

This answer is of course reminiscent of the replies I gave above to other objections.

**The 'sheer nonsense' objection** The last argument Schroeder gives against radical conventionalism is simply that it does not make sense. It is just not the

case that a holding a certain statement fast is evidence of a convention. Schroeder writes:

> A convention is an agreement what to do (not just what to believe) under certain *repeatable* circumstances, in a certain *kind* of situation, not just on one occasion. Hence a referendum, a on-off decision is not a convention. (Schroeder 2017a, p. 95)

Thus, we have conventions on how to calculate, and so on, but not conventions *that* a particular sum is correct. Convention, Schroeder argues, requires generality, and any convention must be applicable to countless particular cases on each occasion, and hence deciding on a case by case basis, as Dummett's reading of Wittgenstein requires, simply means that there is no convention in play. It is, Schroeder concludes, a contradiction in terms.

The account, as presented here, depends on basic constitutive practices, which are, as I argued above, not strictly speaking *conventions*. The account is an instance of conventionalism, however, *because* each truth depends on our agreement, which Schroeder himself agrees is the mark of the conventional. Their generality, in turn, is manifested in that they determine, for each particular case, what counts as as in instance of whatever they are the basic constitutive practice of—e.g. the basic constitutive practice of adding determines for each action whether or not it is adding or not. It might therefore very well be true that the concept of convention that Dummett propounds is incoherent and contradictory, but the account presented here is sufficiently different, all the while being an instance of radical conventionalism, to not succumb to that objection.

These arguments are related to an objection Barry Stroud makes to Dummett's reading (Stroud 1965). For Stroud, it is our form of life that determines that we cannot but accept one way of proceeding when following a rule, and there must be live alternatives for us in order for what we do to count as conventional (ibid., p. 515–516). Stroud cites a famous passage by Wittgenstein:

> I am not saying: if such-and-such facts of nature were different, people would have different concepts (in the sense of a hypothesis). But: if anyone believes that certain concepts are absolutely the correct ones, and that having different ones would mean not realizing something

> that we realize—then let him imagine certain very general facts of nature to be different from what we are used to, and the formation of concepts different from the usual ones will become intelligible to him. (*PI*, xii, p. 195e )

For Stroud, Wittgenstein's examples of different concepts are meant to show that while the formation of different concepts from the usual one is intelligible to us, the concepts themselves, thus formed are not. And, Stroud says, "since the intelligibility of alternative concepts and practices is required by the thesis of radical conventionalism which Dummett ascribes to Wittgenstein" (ibid., p. 515–516), Wittgenstein's examples cannot be meant to show that, and hence Wittgenstein was not a radical conventionalist.

In response to Stroud's point, I would point out, as I have argued above, that it is not necessary for conventionalism that there are different options we can choose from, neither logically nor psychologically. It is enough that correctness or truth in the relevant domain is determined by our practices, and not external reality. Furthermore, our basic constitutive practices *are* arbitrary in Ben-Menahem's sense, since they are not determined by other conventions or external factors.

It is of course true that from a psychological point of view, given our training and form of life, we only see one path available before us, but that fact does not make what we do any less arbitrary. For example, we can imagine, following Tyler Burge, that there is

> a small, isolated, unenterprising linguistic community none of whose members ever heard of anyone's speaking differently. It would not be surprising if there were a few such communities in the world today. It would be amazing if there never had been. Such a community would not know—or perhaps even have reason to believe—that there are humanly possible alternatives to speaking their language. If they were sufficiently ignorant of human learning, they might believe that their principal linguistic regularities were immutably determined by natural law. Yet we have no inclination to deny that their language is conventional. They are simply ignorant or wrong about the nature of their activities. (Burge 1975, p. 250)

Likewise, if we would have had a different training, then we would have different concepts, and in that sense they are arbitrary. It is irrelevant that we cannot imagine it otherwise, when we are engaged in our practice, whatever it is.

### 7.2.3 Putnam's 'consistency'-objection

The last objection I want to consider is Putnam's 'consistency'-objection. This isn't an objection against radical conventionalism as such, but rather a more encompassing objection that includes more moderate readings of Wittgenstein's philosophy.

Putnam starts by contrasting Dummett's reading of Wittgenstein as a radical conventionalist with Stroud's reading mentioned in the previous section. For Putnam, Stroud's reading misses the important philosophical point that Dummett was making, since even if it is true that it is our form of life that fixes which rule we are following, it still implies that mathematical truth is 'up to us':

> The real point is that if either Dummett or Stroud is right, then Wittgenstein is claiming that mathematical truth and necessity arise in us, that it is human nature and forms of life that explain mathematical truth and necessity. If this is right, then it is the greatest philosophical discovery of all time. Even if it is wrong, it is an astounding philosophical claim. If Stroud does not dispute that Wittgenstein advanced this claim—and he does not seem to dispute it—then his interpretation of Wittgenstein is a revision of Dummett's rather than a total rejection of it. (Putnam 1979, p. 425)

So far, I have argued that conventionalism in *general* is indeed the view that Putnam is describing here and calls an 'astounding philosophical claim', namely that mathematical truth 'arises in us', or is up to us—in the sense outlined in Chapter 5, and I have further argued that *radical* conventionalism is the view that our agreement about the particular case, spelled out as basic constitutive practices, is what determines mathematical truth—by determining which mathematical concepts we are in fact employing. On this reading, Stroud is at least in the former camp, that mathematical truth depends on us and our practices.

Putnam's consistency objection, which is meant to target both versions, however, is as follows: Consider some formalisation of number theory, e.g. Peano

Arithmetic. If we suppose that our acceptance of the axioms of PA is just, as Putnam puts it, "the acceptance of a bunch of *meaning determinations*" (ibid., p. 425), they are still required to be *consistent*, and while our form of life might explain why we accept *this* set of axioms, rather than another, it cannot make it true that those axioms are consistent, and therefore there is at least one mathematical fact, namely the consistency of these axioms, however they come to be, which is not explained by Wittgenstein's account, that is to say, the fact that PA is consistent is an objective mathematical fact which does not simply follow empirically from our nature, but is completely independent of it.

Of course, Wittgenstein had much to say about consistency and its status, which he did not think had the philosophical importance most contemporary philosophers think it does, but Putnam correctly points out that even if that might be textually correct and accurately describe Wittgenstein's views, it does not answer the objection. The objection aims to show that there is at least one mathematical fact that is not determined by our human nature and culture, however significant that fact is. Simply rejecting consistency as irrelevant is itself irrelevant.[18]

Putnam thinks that Wittgenstein would have responded to this argument by referring to his rule-following considerations and conclude that mathematical truth simply isn't determined in cases where our practice hasn't manifested itself, and so the statement that PA is consistent may have no truth-value. Putnam is sympathetic to this view and agrees that "even so simple an operation as *modus ponens* is not 'fixed' once and for all by our mental representation of the operation" and that it is "our actual 'unpacking' of the mental representation in action, our *de facto* dispositions which determine what we *mean*..." (ibid., p. 427). However, there are mathematical questions where the number of cases to check is small enough so that we could conceivably check all of them, and there, Wittgenstein's view should predict a determinate answer. That, Putnam thinks, leads to trouble, even if Wittgenstein is right about how our *de facto* dispositions determine what we mean.

Now consider the concept '$x$ follows from $y$ and $z$ by *modus ponens*'. There are, Putnam purports, two possible scenarios as to how our dispositions might fix the

18. For a discussion and overview of Wittgenstein's remarks concerning contradictions and inconsistency, see **Matthiassson2020le**.

meaning of a symbol referring to this concept where the number of cases is small enough, say involving proofs with fewer than $10^{20}$ symbols.

Scenario (1) is as follows: we are given a proof in PA with fewer than $10^{20}$ symbols and check it by going down line by line and verifying each line by making sure that it is either an axiom or follows from one of the previous lines by *modus ponens*. If the last line is $1 = 0$, we say that Peano Arithmetic has turned out to be inconsistent. In scenario (2) we proceed as in scenario (1), except that when we reach the last line, we *modify* the concept '*x* follows from *y* and *z* by *modus ponens*' so that the last line does not follow, instead of accepting the inconsistency.

These scenarios are both logically possible, Putnam thinks. However, only in scenario (2) does Putnam thinks that the fact that PA is consistent 'arise in us'—"'be explained by our nature (our dispositions) in a clear sense" (Putnam 1979, p. 427) and that scenario does not obtain. We are, in fact, in scenario (1), and there, the $10^{20}$-consistency of PA is, in Putnam's estimation "is still not an artifact of this dispositionally fixed interpretation" (ibid., p. 427). It is not, in this case, Putnam think, a truth that is fully explained by human nature that PA is thus consistent and thus he concludes that Wittgenstein's view that mathematical facts are not objective facts cannot be right.

How could we respond to this argument? First of all, I agree with Putnam that we are in scenario (1), and not in scenario (2)—which looks as if it is supposed to correspond to Dummett's version of radical conventionalism—but it does not follow that radical conventionalism is therefore false, since the version defended here can in fact accommodate scenario (1).[19]

Consider a much simpler example. Suppose a formal system we are considering has among its axioms $q$, $q \to p$ and $q \to \neg p$, as well as at least modus ponens and conjunction introduction as inference rules. This system is trivially inconsistent

19. And it should be noted that if it can, that is a further blow to more moderate readings of Wittgenstein, since Putnam's argument would still bite against them.

and an elementary proof runs as follows:

| | | |
|---|---|---|
| (1) | $q$ | Axiom. |
| (2) | $q \to p$ | Axiom. |
| (3) | $q \to \neg p$ | Axiom. |
| (4) | $p$ | *Modus ponens* from (1) and (2). |
| (5) | $\neg p$ | *Modus ponens* from (1) and (3). |
| (6) | $p \wedge \neg p$ | *Conjunction introduction* from (4) and (5). |

Now recall that on the radical conventionalist account developed here, that we are in fact using the rules *modus ponens* and *conjunction introduction*, and not some other deviant rules, is fixed by our basic constitutive practices of using the terms 'modus ponens' and 'conjunction introduction', and there, our agreement about the particular case is constitutive of which rule we are in fact using: if our stable dispositions to judgement were stable where line (4) does not follow from (1) and (2), we'd be using some other rule than *modus ponens*. That simply follows from the nature of the account.

It is therefore true in Putnam's sense, that *given the meaning* of 'modus ponens' and 'conjunction introduction', it is an objective truth that $p \wedge \neg p$ can be derived in the system, and hence an objective truth that it is inconsistent. The meanings of the terms 'modus ponens' and 'conjunction introduction', however, are determined by our basic constitutive practice of using them, and so the correctness conditions of these practices are determined by our stable dispositions to judgement, including the particular case. The same goes for the concept *inconsistent*—that the system is inconsistent if $p \wedge \neg p$ can be derived is what it *means* to be inconsistent, and the meaning of 'inconsistent' is likewise determined by the basic constitutive practice of using the term, where our agreement about this particular case we are considering now is *also* constitutive of its meaning. That $p \wedge \neg p$ can be derived in the system is just what it *is* for it to be inconsistent.

Hence, since nothing else than the meaning of the terms, determined by our practice, makes a contribution to the truth of the statement that the system is inconsistent, and that those meanings are constituted by our agreement to stable

judgement about the particular case, its truth is directly determined by our mathematical practice, just as radical conventionalism would predict. It was never the radical conventionalist's claim that mathematical truth is not objective, just that they are linguistic truths in the sense outlined above, and that certainly seems to be the case here.

Therefore, Putnam's objection fails in this case, and since our example is analogous to scenario (1), Putnam's objection fails in general. It should, however, be noted that nowhere does my counter-argument rely on placing an upper bound on the size of the proof being examined, and hence the result is even stronger than the one Putnam argues against.

# Conclusion

In the first part of this thesis, I developed a game-theoretic solution to the rule-following paradox inspired by Wittgenstein's discussion of rule-following as practice in the *Philosophical Investigations*. I argued that this solution is able to account for the correctness conditions of rule-following and meaning, for indefinitely many cases and without circularity by defining correctness as a point on a correlated equilibrium of a basic constitutive practice. The solution places heavy emphasis on training, agreement in judgement and our form of life.

In the second part, I used this solution to deflect criticism of Dummett's reading of Wittgenstein as a radical conventionalist. I argued that decision is not a fundamental part of what we should take radical conventionalism to be and that it should rather be seen through the contrast with more moderate forms of conventionalism, whereby an unreduced notion of consequence is appealed to in order to move from stipulated truths to further, more remote truths. On this view, radical conventionalism is conventionalism all the way down: there is no external criterion at all for the correctness of each step in a mathematical proof except our own practice.

By identifying mathematical correctness with correctness in basic constitutive practices the view is able to avoid the problems that dogged moderate forms of conventionalism, e.g. Quine's regress problem, as the game-theoretic structure of such practices is able to define correctness for indefinitely many cases without appealing to anything outside itself. I then used that solution to respond to common arguments against radical conventionalism. I concluded that it remains a viable view in the philosophy of mathematics.

# Appendix A

# Formal details

In this appendix, I outline the formal details which the solution to the rule-following paradox given in Chapter 4 depends. The appendix is in two parts. The first part contains the game-theoretic framework I borrow from Peter Vanderschraaf's book (Vanderschraaf 2018, Chapter 2 and Appendix 1) and the second part contains my own extension of Vanderschraaf's concept of an equilibrium path.

## A.1 A fraction of Peter Vanderschraaf's theory of convention

We let $(\Gamma_t)$ be a sequence of games $\Gamma_1, \Gamma_2, \ldots$ indexed by a *time period* $t$ (where each $\Gamma_i$ is some interaction like the coordination problems discussed in Chapter 4). We say that such a sequence $(\Gamma_t)$ is a supergame. Let $N = \{1, 2, \ldots\}$ be a community of agents such that at each successive time period $t$, a finite and non-empty subset $N_t \subseteq N$ is chosen to play a game $\Gamma_t$.

To each agent $i$, we associate a set of strategies $A_i = \{a_{il}, \ldots, a_{im}\}$. If $i$ is not chosen to play at a given time period $t$, we say that $i$ plays the empty move, denoted by $\underline{a}$. At each time period $t \geq 1$ every agent makes a move $s_{ti} \in A_i \cup \{\underline{a}\}$. An act profile $\boldsymbol{s}_t = \{s_{1t}, s_{2t}, \ldots\}$ is the set of moves of all agents at $t$. At each act profile $\boldsymbol{s}_t$ an agent $i$ receives a pay-off $u_i(\boldsymbol{s}_t)$ where $u_i(\boldsymbol{s}_t) = 0$ if $s_{it} = \underline{a}$. If there are only two agents playing, then this definition simply describes a pay-off matrix of a game, where each row describes the possible moves of agent $i$ and each column

the moves of $j$. Each cell in the matrix represents the pay-offs of each agent given an act profile.

To each supergame $(\Gamma_t)$ we associate a set $\Omega$ of possible worlds where one of these is the actual world at a given stage $t$, denoted by $\omega(t)$. Each agent $i$ has a *private information partition* $\mathcal{H}_i$ of $\Omega$ and a *conjecture* $\mu_i(\cdot)$ over $\mathcal{H}_i$. Together, these represent $i$'s beliefs about what the actual world is and how likely $i$ thinks each possibility is. At each period $t \geq 1$, we say that the set $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{t-1}\}$ represents the *ex ante history* of interactions. A full description of $\omega(t)$ includes the all the information relevant to $i$'s decisions, including who the other agents are, what the history is, the agents' private information partitions, conjectures, etc.

In addition, let $f_i : \Omega \to A_i$ be a function that represents the *individual strategies* for each agent $i$ and is constant over each cell of $\mathcal{H}_i$. This tells us which action $i$ takes, given his beliefs about what the actual world is. This function being constant over the cells of $\mathcal{H}_i$ means that if any two worlds $\omega_k$ and $\omega_l$ are in the same cell of a partition, then $f_i(\omega_k) = f_i(\omega_l)$. Given the individual strategies of the agents, we can define a strategy profile $\boldsymbol{f} = (f_1, f_2, \ldots)$ over all the agents. $\boldsymbol{f}$ tells us which action profile results from the actions of each agent in each possible world.

We also define a discount factor $\delta_i \in (0, 1)$. This can either be understood as representing how much the agents value their payoffs in the next period compared to the current one (the intuition is that it is reasonable to prefer one pound today rather than one tomorrow, if one is not sure one will play tomorrow) or the likelihood that the game continues beyond the next period. The discount factor has the convenient mathematical property of keeping the pay-offs of a player finite over an infinite sequence of games. Otherwise it will not really play any role for us, and it is strictly speaking not necessary.

Finally, we define agent's $i$ expected pay-off at a world , given a strategy profile $\boldsymbol{f} = (f_1, f_2, \ldots)$ as

$$E_i(u_i(\boldsymbol{f})) = \sum_{t=1, \omega \in \Omega}^{\infty} u_i(\boldsymbol{f}(\omega)) \times \mu_i[\omega(t) = \omega] \times \delta_i^t$$

That is to say, $i$'s expected utility is the sum of $i$'s pay-offs at $t$ given a strategy profile $\boldsymbol{f}$ and a world $\omega$, multiplied by $i$'s subjective probability of $\omega$ being the actual world, multiplied by $i$'s discount factor. We say that $\boldsymbol{f}$ is a *correlated equilibrium* of the

supergame $(\Gamma_t)$ if for each $i \in N$ and any strategy $h_i \neq f_i$,

$$E_i(u_i(\boldsymbol{f})) \geq E_i(u_i(h_i,\boldsymbol{f}_{-i})).$$

The notation $(h_i,\boldsymbol{f}_{-i})$ refers to the sequence that has $f_i$ replaced by $h_i$ in $\boldsymbol{f}$, and so the intended interpretation of this definition is that if $i$ unilaterally deviates from $\boldsymbol{f}$, then $i$'s pay-off is strictly lower than it would have been if $i$ had played $f_i$. If we suppose that the conjectures of the agents are probabilistically independent, then this definition coincides with that of the Nash equilibrium (see Vanderschraaf 2018, 72).

## A.2 First and second-order equilibrium paths

A sequence of act profiles $\boldsymbol{s}_t = \{s_{1t}, s_{2t}, \ldots\}$ is an *equilibrium path* of $\boldsymbol{f}$, if for all $t$, every $s_{it} \in \boldsymbol{s}_t$ is performed according to some strategy $f_i$ of a strategy profile $\boldsymbol{f}$ which is in equilibrium. If $\boldsymbol{f}$ is an equilibrium of a supergame $(\Gamma_t)$, its equilibrium paths are equilibrium paths of $(\Gamma_t)$. For example, in the meeting game considered in the Chapter 4, and assuming that $N = \{R, C\}$ and $A_i = \{1, 2, 3\}$, the sequence of act profiles $\boldsymbol{s}_1 = \{1_{R1}, 1_{C1}\}, \boldsymbol{s}_2 = \{1_{R2}, 1_{C2}\} \ldots$, represents an equilibrium path of the meeting game where both agents always go to the place denoted by 1. This game has infinitely many equilibrium paths.

Now, let $(\Gamma)_* = (\Gamma_t)_1, (\Gamma_t)_2 \ldots$ be a sequence of supergames. A selection of one equilibrium path from each supergame in $(\Gamma)_*$ is a *second-order equilibrium path* of $(\Gamma)_*$.

If $(\Gamma)_*$ represents a practice of using a term denoting a concept, for instance the practice of using the symbol '+', then each second-order equilibrium path represents one possible interpretation of the term. For instance, if each $(\Gamma_t)_i$ represents a use of '+' of the form '$n + m = x$', where $x$ represents the answer the agents can give to such an addition problem, then a selection of equilibrium paths represents the concept of *quaddition*, if we choose the paths corresponding to addition up to $n = 57$ and $\boldsymbol{s}_t = \{5, 5, \ldots\}$ for any game afterwards.

# Bibliography


Aumann, Robert J. 1974. "Subjectivity and Correlation in Randomized Strategies." *Journal of Mathematical Economics* 1:67–96.

———. 1987. "Correlated Equilibrium as an Expression of Bayesian Rationality." *Econometrica* 55 (1): 1–18.

Bangu, Sorin. 2012a. "Later Wittgenstein and the Genealogy of Mathematical Necessity." In *Wittgenstein and Naturalism,* edited by Kevin M. Cahill and Thomas Raleigh.

———. 2012b. "Later Wittgenstein's Philosophy of Mathematics." In *Internet Encyclopedia of Philosophy,* edited by J. Feiser and B. Dowden.

———. 2012c. "Wynn Experiments and the Later Wittgenstein Philosophy of Mathematics." *Iyyun* 61:219–240.

Ben-Menahem, Yemima. 1998. "Explanation and Description: Wittgenstein on Convention." *Synthese* 115 (1): 99–130.

———. 2006. *Conventionalism: From Poincare to Quine.* Cambridge: Cambridge University Press.

Benacerraf, Paul. 1973. "Mathematical Truth." *Journal of Philosophy* 70 (19): 661–679.

Blackburn, Simon. 1984a. *Spreading the Word: Groundings in the Philosophy of Language.* Oxford: Clarendon.

———. 1984b. "The Individual Strikes Back." *Synthese* 58 (March): 281–302.

Bloor, David. 1997. *Wittgenstein, Rules and Institutions.* London: Routledge.

Boghossian, Paul. 1989. "The Rule-Following Considerations." *Mind* 98 (392): 507–549.

———. 1996. "Analyticity Reconsidered." *Noûs* 30 (3): 360–391.

Burge, Tyler. 1975. "On Knowledge and Convention." *The Philosophical Review* 84 (2): 249–255.

———. 1979. "Individualism and the Mental." *Midwest Studies in Philosophy* 4 (1): 73–122.

———. 1986. "Individualism and Psychology." *The Philosophical Review* 95 (January): 3–45.

Byrne, Alex. 1996. "On Misinterpreting Kripke's Wittgenstein." *Philosophy and Phenomenological Research* 56 (2): 339–343.

Dummett, Michael. 1959. "Wittgenstein's Philosophy of Mathematics." *The Philosophical Review* 68 (3): 324–348.

———. 1978. *Truth and Other Enigmas.* Cambridge: Harvard University Press.

———. 1986. "A Nice Derangement of Epitaphs: Some Comments on Davidson and Hacking." In *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson,* edited by Ernest LePore. Cambridge: Blackwell.

———. 1993. "Wittgenstein on Necessity: Some Reflections." In *The Seas of Language,* 446–461. Oxford: Oxford University Press.

Finkelstein, David H. 2000. In *The New Wittgenstein,* edited by Alice Crary and Rupert Read, 83–100. London: Routledge.

Finn, Suki. 2019. "The Adoption Problem and Anti-Exceptionalism About Logic." *Australasian Journal of Logic* 16 (7): 231.

Fogelin, Robert J. 1995. *Wittgenstein.* London: Routledge.

———. 2009. *Taking Wittgenstein at his Word.* Princeton: Princeton University Press.

Gerrard, Steven. 1991. "Wittgenstein's Philosophies of Mathematics." *Synthese* 87 (1): 125–142.

———. 2018. "A Philosophy of Mathematics Between Two Camps." In *The Cambridge Companion to Wittgenstein,* Second edition, edited by David G. Stern Hans Sluga. Cambridge: Cambridge University Press.

Gintis, Herbert. 2009. *The Bounds of Reason.* Game Theory and the Unification of the Behavioural Sciences. Princeton: Princeton University Press.

Glock, Hans-Johann. 1996. "Necessity and Normativity." In *The Cambridge Companion to Wittgenstein,* edited by Hans D. Sluga and David G. Stern, 198–225. Cambridge: Cambridge University Press.

———. 2003. "The Linguistic Doctrine Revisited." *Grazer Philosophische Studien* 66 (1): 143.

———. 2008. "Necessity and Language: In Defence of Conventionalism." *Philosophical Investigations* 31 (1): 24–47.

Glüer, Kathrin, and Peter Pagin. 1999. "Rules of Meaning and Practical Reasoning." *Synthese* 117 (2): 207–227.

Glüer, Kathrin, and Åsa Wikforss. 2009. "Against Content Normativity." *Mind* 118 (469): 31–70.

———. 2010. "Es Braucht Die Regel Nicht: Wittgenstein on Rules and Meaning." In *The Later Wittgenstein on Language,* edited by Daniel Whiting. Basingstoke: Palgrave-Macmillan.

———. 2018. "The Normativity of Meaning and Content." In *The Stanford Encyclopedia of Philosophy,* Spring 2018, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Goldfarb, Warren. 2012. "Rule-Following Revisited." In *Wittgenstein and the Philosophy of Mind,* edited by Jonathan Ellis and Daniel Guevara. Oxford: Oxford University Press.

Guala, Francesco, and Frank Hindriks. 2014. "A Unified Social Ontology." *The Philosophical Quarterly* 65 (259): 177–201.

Hacker, P. M. S., and G. P. Baker. 1984a. "On Misunderstanding Wittgenstein: Kripke's Private Language Argument." *Synthese* 58 (3): 407–450.

———. 1984b. *Scepticism, Rules and Language.* Oxford: Basil Blackwell.

———. 2009. *Wittgenstein: Rules, Grammar and Necessity.* Second Edition. An Analytical Commentary on the Philosophical Investigations. Oxford: Wiley-Blackwell.

Hacking, Ian. 2014. *Why Is There Philosophy of Mathematics At All?* Cambridge: Cambridge University Press.

Hart, W D. 1991. "Benacerraf's Dilemma." *Critica Revista Hispanoamericana de Filosofia* 23 (68): 87–103.

Hattiangadi, Anandi. 2006. "Is Meaning Normative?" *Mind and Language* 21 (2): 220–240.

———. 2007. *Oughts and Thoughts: Rule-Following and the Normativity of Content.* Oxford: Oxford University Press.

Hilmy, S. Stephen. 1987. *The Later Wittgenstein: The Emergence of a New Philosophical Method.* New York: Blackwell.

Hintikka, Jaakko. 1989. "Rules, Games and Experiences: Wittgenstein's Discussion of Rule-Following in the Light of His Development." *Revue Internationale de Philosophie* 43 (169): 279–297.

Kripke, Saul. 1982. *Wittgenstein on Rules and Private Language.* Oxford: Basil Blackwell.

Kusch, Martin. 2002. *Knowledge by Agreement: The Programme of Communitarian Epistemology.* Oxford: Oxford University Press.

———. 2006. *A Sceptical Guide to Meaning and Rules: Defending Kripke's Wittgenstein.* Chesham: Acumen.

Lewis, C. I. 1946. *An Analysis of Knowledge and Valuation.* La Salle: The Open Court Publishing Company.

Lewis, David. 1969. *Convention: A Philosophical Study.* Cambridge: Harvard University Press.

Lewy, Casimir. 1976. *Meaning and Modality.* Cambridge: Cambridge University Press.

Linnebo, Øystein. 2017. "Platonism in the Philosophy of Mathematics." In *The Stanford Encyclopedia of Philosophy,* Summer 2017, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Maddy, Penelope. 1990. *Realism in Mathematics.* Oxford: Oxford University Press.

Marcus, Russell. 2015. *Autonomy Platonism and the Indispensability Argument.* Lanham: Lexington Books.

Marmor, Andrei. 2009. *Social Conventions: From Language to Law.* Princeton: Princeton University Press.

Matthíasson, Ásgeir Berg. Forthcoming. "Contradictions and falling bridges: What was Wittgenstein's reply to Turing?" *British Journal for the History of Philosophy.*

McDowell, John. 1984. "Wittgenstein on Following a Rule." *Synthese* 58 (March): 325–364.

———. 1992. "Meaning and Intentionality in Wittgenstein's Later Philosophy." *Midwest Studies in Philosophy* 17 (1): 40–52.

McEvoy, Mark. 2012. "Platonism and the 'Epistemic Role Puzzle'." *Philosophia Mathematica* 20 (3): 289–304.

Midgley, G. C. J. 1959. "XIV—Linguistic Rules." *Proceedings of the Aristotelian Society* 59 (1): 271–290.

Millar, Alan. 2002. "The Normativity of Meaning." *Royal Institute of Philosophy Supplement* 51:57–73.

Miller, Alexander. 2002. "Introduction." In *Rule-Following and Meaning,* edited by Alexander Miller and Crispin Wright. Montreal & Kingston: McGill-Queen's University Press.

Monk, Ray. 1991. *Wittgenstein: The Duty of Genius.* New York: Penguin Books.

Padró, Romina. 2015. "What the Tortoise Said to Kripke: the Adoption Problem and the Epistemology of Logic." PhD diss., City University of New York.

Potter, Michael. 2008. *Wittgenstein's Notes on Logic.* Oxford: Oxford University Press.

Putnam, Hilary. 1979. "Analyticity and Apriority: Beyond Wittgenstein and Quine." *Midwest Studies in Philosophy* 1 (4): 423–441.

Quine, Willard Van Orman. 1966. "Truth by Convention." In *The Ways of Paradox and Other Essays,* 70–99. New York: Random House.

———. 1968. "Ontological Relativity." *Journal of Philosophy* 65 (7): 185–212.

Rawls, John. 1955. "Two Concepts of Rules." *The Philosophical Review* 64 (1): 3–32.

Reiland, Indrek. 2019. "Constitutive Rules: Games, Language, and Assertion." *Philosophy and Phenomenological Research.*

Resnik, Michael. 1997. *Mathematics as a Science of Patterns.* New York: Oxford University Press.

Resnik, Michael D. 1981. "Mathematics as a Science of Patterns: Ontology and Reference." *Noûs* 15 (4): 529–550.

Rhees, Rush. 1958. "Introduction." In *The Blue and Brown Books: Preliminary Studies for the "Philosophical Investigations",* by Ludwig Wittgenstein. Oxford: Basil Blackwell.

Ruben, David-Hillel. 1997. "John Searle's The Construction of Social Reality." *Philosophy and Phenomenological Research* 57 (2): 443–447.

Schroeder, Severin. 2014. "Mathematical Propositions as Rules of Grammar." *Grazer Philosophische Studien* 89:21–36.

———. 2017a. "On some standard objections to mathematical conventionalism." *Belgrade Philosophical Annual,* no. 30: 83–98.

———. 2017b. "Wittgenstein on Grammar and Grammatical Statements." In *A Companion to Wittgenstein,* edited by Hans-Johann Glock and John Hyman, 252–268. Oxford: Wiley-Blackwell.

Searle, John. 1969. *Speech Acts: An Essay in the Philosophy of Language.* Cambridge: Cambridge University Press.

Sellars, Wilfrid. 1997. *Empiricism and the Philosophy of Mind.* Cambridge: Harvard University Press.

Shanker, Stuart. 2005. *Wittgenstein and the Turning Point in the Philosophy of Mathematics.* New York: Routledge.

Shapiro, Stewart. 1997. *Philosophy of Mathematics: Structure and Ontology.* Oxford: Oxford University Press.

Shogenji, Tomoji. 1993. "Modest Scepticism About Rule-Following." *Australasian Journal of Philosophy* 71 (4): 486–500.

Sider, Theodore. 2003. "Reductive Theories of Modality." In *The Oxford Handbook of Metaphysics,* edited by Michael J. Loux and Dean W. Zimmerman, 180–208. Oxford: Oxford University Press.

Sillari, Giacomo. 2013. "Rule-Following as Coordination: A Game-Theoretic Approach." *Synthese* 190 (5): 871–890.

Soames, Scott. 1998. "Facts, Truth Conditions, and the Skeptical Solution to the Rule-Following Paradox." *Philosophical Perspectives* 12 (S12): 313–348.

Steiner, Mark. 2000. "Mathematical Intuition and Physical Intuition in Wittgenstein's Later Philosophy." *Synthese* 125 (3): 333–340.

———. 2009. "Empirical Regularities in Wittgenstein's Philosophy of Mathematics." *Philosophia Mathematica* 17 (1): 1–34.

Stern, David G. 2018. "Wittgenstein in the 1930s." In *The Cambridge Companion to Wittgenstein,* Second edition, edited by David G. Stern Hans Sluga. Cambridge: Cambridge University Press.

Stroud, Barry. 1965. "Wittgenstein and Logical Necessity." *The Philosophical Review* 74 (October): 504–518.

Topey, Brett. 2019. "Linguistic Convention and Worldly Fact: Prospects for a Naturalist Theory of the a Priori." *Philosophical Studies* 176 (7): 1725–1752.

Unnsteinsson, Elmar Geir. 2016. "Wittgenstein as a Gricean Intentionalist." *British Journal for the History of Philosophy* 24 (1): 155–172.

Vanderschraaf, Peter. 1995. "Convention as Correlated Equilibrium." *Erkenntnis* 42 (1): 65–87.

———. 1998. "Knowledge, Equilibrium and Convention." *Erkenntnis* 49 (3): 337–369.

———. 2018. *Strategic Justice: Convention and Problems of Balancing Divergent Interests.* Oxford: Oxford University Press.

Verheggen, Claudine. 2003. "Wittgenstein's Rule-Following Paradox and the Objectivity of Meaning." *Philosophical Investigations* 26 (4): 285–310.

———. 2011. "Semantic Normativity and Naturalism." *Logique Et Analyse* 54 (216): 553–567.

Vision, Gerald. 2005. "The Truth about Philosophical Investigations I §§134–137." *Philosophical Investigations* 28 (2): 159–176.

Warren, Jared. 2015. "The Possibility of Truth by Convention." *The Philosophical Quarterly* 65 (258): 84–93.

———. 2016. "Revisiting Quine on Truth by Convention." *Journal of Philosophical Logic* 46 (2): 1–21.

———. 2018. "Killing Kripkenstein's Monster." *Noûs.*

Whiting, Daniel. 2007. "The Normativity of Meaning Defended." *Analysis* 67 (2): 133–140.

———. 2016. "What is the Normativity of Meaning?" *Inquiry : An Interdisciplinary Journal of Philosophy* 59 (3): 219–238.

Wikforss, Åsa. 2001. "Semantic Normativity." *Philosophical Studies* 102 (2): 203–226.

———. 2016. "Davidson and Wittgenstein – a Homeric Struggle?" In *Wittgenstein and Davidson on Thought, Language, and Action,* edited by Claudine Verheggen. Cambridge: Cambridge University Press.

Williams, Meredith. 1999. *Wittgenstein, Mind, and Meaning: Toward a Social Conception of Mind.* London: Routledge.

Wilson, George. 1998. "Semantic Realism and Kripke's Wittgenstein." *Philosophy and Phenomenological Research* 58 (1): 99–122.

Wright, Crispin. 2001a. "Does *Philosophical Investigations* §§258–60 Suggest a Cogent Argument against Private Language?" In Wright 2001d, 223–279.

———. 2001b. "Excerpts from a Critical Study of Colin McGinn's *Wittgenstein on Meaning*." In Wright 2001d, 143–169.

———. 2001c. "Kripke's Account of the Argument against Private Language." In Wright 2001d, 91–115.

———. 2001d. *Rails to Infinity: Essays on Themes from Wittgenstein's Philosophical Investigations.* Cambridge: Harvard University Press.

———. 2001e. "Wittgenstein's Rule-Following Considerations and the Central Project of Theoretical Linguistics." In Wright 2001d, 170–213.

———. 2002. "Meaning and Intention as Judgement Dependent." In *Rule-Following and Meaning,* edited by Alexander Miller and Crispin Wright. Montreal & Kingston: McGill-Queen's University Press.

———. 2007. "Rule-Following Without Reasons: Wittgenstein's Quietism and the Constitutive Question." *Ratio* 20 (4): 481–502.

Wright, Crispin. 2012. "Replies Part I: The Rule-Following Considerations and the Normativity of Meaning." In *Mind, Meaning, and Knowledge: Themes From the Philosophy of Crispin Wright,* edited by Crispin Wright and Annalisa Coliva, 201–219. Oxford: Oxford University Press.