Preprint of a paper forthcoming in *Inquiry*. Please only cite published version when it becomes available.

Rules as constitutive practices defined by correlated equilibria

Ásgeir Berg Matthíasson
asgeirberg@hi.is

> "Our language does have a fairly determined interpretation (a Moorean fact!) so there must be some constraint not created ex nihilo by our stipulation."
>
> —David Lewis

In his famous interpretation of Wittgenstein's discussion of rule-following, Kripke develops the following paradox. 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' that 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 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.

In this paper, I give a solution to this paradox in terms of what I will call *basic constitutive practices.* I will argue that by defining such practices in game-theoretic terms we can give a satisfactory reply to Kripke’s sceptic that defines the correctness conditions of our most basic concepts as those actions that lie on the correlated equilibrium of such a practice. The resulting picture of language will preserve the objectivity and correctness conditions of meaning, but do away with guidance—all the while motivating why it would seem that meaning plays this role, without actually needing it. A major benefit of the account is that it can give replies to important and well-known objections to communitarian accounts of rule-following.

The paper is in three main parts. I will start by giving an outline of the problem, saying a bit more about what I believe our desiderata for a theory of meaning should be, next I explain my notion of basic constitutive practice and how we can use game-theory to give it the necessary structure to solve the problem, as well as replying to some worries that might be raised about the account. I end by saying a few words on rule-following more generally, and I argue that by accepting the account presented here, we have a strong reason to reject the idea that language is rule-governed, since certain difficult puzzles about the relationship between meaning and rules are dissolved.

1 An outline of the problem

The problem just outlined 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.

Although Kripke gives this problem in the form of a problem about rule-following, it really is about the determination of meaning or content by finite means, and it is helpful, I believe, to look at it through the lens of how we might learn a language: 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.[1]

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

Kripke's argument proceeds by elimination. He 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 follows a particular algorithm, that meaning is a primitive and cannot be given further analysis, platonism, and many others. Most of Kripke's attention, however, 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

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

not their quum. His most important argument against this kind of account concerns the normativity of meaning: that if one means something by a term, then one *should* or *ought* to use it in a specific way. Kripke writes:

> 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' [. . .] The relation of meaning and intention to future action is *normative*, not descriptive. (Kripke 1982, p. 37)

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

There is little consensus in the literature what this objection amounts to, and in what follows, I will adopt a very weak understanding of the normativity of meaning and simply assume that was is at stake is the 'platitudinous' idea (Wright 2001, 276) that our use of words has conditions of correct use—that when we use words in any type of linguistic intercourse, some uses of that word will be right and some wrong.[2] On the way I'm construing the problem above, 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 picking out one sameness relation from past exemplars to future cases as correct—and thereby setting up 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. Mere dispositions don't seem to be able to fit the bill, since they are not the kind of thing that could pick out one sameness relation among many as being correct.

Accordingly, the solution I propose will focus on the possibility of applying a word correctly or incorrectly to an object—e.g. that if S means *blue* by his utterance of 'blue', then his application of this word to an object is either correct or incorrect. There are trivial ways to meet this condition, however, and we should also rule out accounts that judge *any* utterance S might ever make as being correct—namely, relative to *some* concept. In other words,

[2] Cf. Boghossian 1989, p. 513.

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*.[3]

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 has conditions of correct use, then, is to say that there is some standard that S adheres to when he applies a word to an object—that there is a way of using words correctly or incorrectly.[4]

As I stated above, there is no consensus in the literature how to read Kripke's argument from normativity, and many philosophers give it a stronger reading than the one I have given it. Here, I'm therefore tentatively siding with those who we can call the anti-normativists about meaning (Hattiangadi 2006, 2007; Wikforss 2001) against the normativists (McDowell 1984; Boghossian 1989; Whiting 2007, 2009, 2016; Peregrin 2012) in how I'm stating the problem (although my way is not the anti-normativists way either). However, solving the weaker problem, of how meaning can have conditions of correct application, is difficult enough, and it does not follow from my

[3] See Wikforss 2001, for a similar discussion.

[4] This corresponds to the distinction Anandi Hattiangadi makes between the *normative* proper and the *norm-relative* (Hattiangadi 2007). Meaning on this view is not normative, but norm-relative: 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). See also Hattiangadi 2006; Glüer and Pagin 1999.

account that meaning is *not* normative in the stronger sense that the normativists claim it is—and it might very well be consistent with a stronger solution to the paradox than the one I am offering here. This should not be controversial, since both sides agree that meaning *at least* has conditions of correct application in the sense outlined, the problem I aim to solve.

On the other hand, as we will see, there is a way to interpret the account so that the normativity of meaning can 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). This notion of normativity is not the one most normativists are after, however. In the example above, S *is* adhering to such a standard, despite his mistake.

2 Rules and meaning as constitutive practices defined by correlated equilibria

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.[5] 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.

I will first explain the notion of constitutive practice that I'm working with, then give the formal 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. After I have given my account below, I will discuss the reasons for this.

2.1 Constitutive practices

Searle's distinction between *regulative rules* and *constitutive rules* is a familiar one (Searle 1969, 33–42). The former kind of rules regulate some activity

[5] See Vanderschraaf 2018.

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. A paradigm example of regulative rules are the rules of traffic. Traffic can and has existed without any rules and explaining what it is does not require a reference to any rules. 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.[6]

Thus for Searle, constitutive rules can be said to "create and define new forms of behaviour", as he puts it (Searle 1969, 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. The important idea for us here is therefore that constitutive rules *constitute* some practice—make that very practice possible. The rules of football, for instance, make it possible play football and 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.[7] If it turns out, however, as I will argue, that rules are constitutive practices themselves, this result of the distinction will not be a problem to be accounted for, but rather an expected feature.

Furthermore, since I want to explain rules as constitutive practices, I cannot without circularity rely on a specification of what that is which essentially relies on rules. I will take my cue from Rawls instead, who suggests

[6] Cf. Glüer and Pagin 1999, 221.

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

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, 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.

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 particular 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. Also note that I do not require that *rules* provide this stagesetting, but am merely using them as examples of what stagesetting could be. This distinction will turn out to be crucial, since if *all* stagesetting comes in the form of rules, the account would be circular.

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. Unlike Rawls, however, I do *not* require that stagesetting takes the form of a constitutive rule, but will appeal to the *structure* of the practice to provide the stagesetting that I require. Basic constitutive practices are therefore not rule-governed, as correctness in a basic constitutive practice *does not* require a prior constitutive rule to be followed, but is defined by the structure of the practice.[8] Therefore, constitutive practices, such as chess, *can* be rule-governed, but basic constitutive practice are *not*, since they appeal to their own structure for correctness, and not further rules.

The idea that I will develop in the rest of the paper is that rules (and meaning) are basic constitutive practices: to follow a 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, which in turn appeals to its own structure to determine correctness. 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 again 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 paper is that meaning and constitutive rules are *both* explained by what I call constitutive practices and that it is the structure of such practices that provides the necessary stagesetting.

What is left then, is to account for these 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.

[8] This notion of "stagesetting" is *broader* than e.g. that of Glüer and Pagin. They write:

> However, what counts as a stagesetting cannot be simply physical, or physically specifiable, circumstances. For, again, 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. And this does indeed indicate what is essential to the idea of stage-setting: that the relevant rules are in force. When we decide to start a game of soccer, what we decide is that the rules of soccer shall start to apply, i.e., be in force for us. We decide what to count as the field, the goal posts, the teams etc., and then, as we proceed to play, our actions are to be evaluated by the rules. (1999: 221)

I agree that stage-setting cannot be physical, or physically specifiable, circumstances, but as we will see, stage-setting does not need to be *rules* that are in force. In basic cases, that is not so.

2.2 Preliminaries: the meeting game

The most fruitful approach to explain how basic constitutive practices can have the features that we require—to provide the stagesetting for itself, as it were, seems to me to be a game-theoretic approach—to impose some structure on the practice itself and use that structure as our criterion of what it is to take part in that practice. Accordingly, my intention is to analyse basic constitutive practices in terms of coordination problems, using a recent account of convention by Peter Vanderschraaf as my framework (Vanderschraaf 2018).[9]

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 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.[10]

Let's start with a simple example of a coordination game, taken from Lewis (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

[9] 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, their theory is not meant to solve the rule-following paradox and the rules that they are concerned with are not basic in the sense that has concerned me here.

[10] In order to reduce the burden on the reader, I will explain my account with a minimal amount of formal details. Instead, I refer the reader to Vanderscrhaaf's book (Vanderschraaf 2018) to get the full formal story.

[11] "R" for "Row" and "'C" for "'Column".

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 with the following matrix (where the first number in the ordered pairs represents the pay-off of R and the second that of C): 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 any agent would unilaterally change his move, he would be worse off.

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

Figure 1: Meeting game

If we suppose that the agents keep meeting each other regularly, we can define a sequence of games $(\Gamma_t) = \Gamma_1, \Gamma_2, \dots$ where each Γ_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 Γ. 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 is not until the agents have met for the first time that f_i settles which action the agents will take.

We say that 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 (Γ_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 (Γ), 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 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 will turn out to be vitally important for the solution on offer.[12]

2.3 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

[12] There is not space here to explicate Sillari's game-theoretic solution to the paradox here, but by relying on Lewis's theory of convention, his solution does not have this property, since Lewis's theory does not have the resources to distinguish between the strategy and the actual acts performed. I hope it should be clear through out the paper why that turns out to be fatal.

case, the practice of using the symbol in the language has this structure in basic cases. This structure is the stagesetting 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 they 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.[13]

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 form the disposition to reply '125' and not '5' when asked what the sum of 57 and 68 is and likewise form the belief that giving this reply is what adding these two numbers is.

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 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 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

	'5'	'125'
'5'	$1, 1$	$0, 0$
'125'	$0, 0$	$1, 1$

Figure 2: Simplified '+' game for two agents

[13] One might think that this assumption is assuming that the problem is solved, and therefore too substantial. For discussion of why this is not so, see below.

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 are formed after similar training), we call this a *correlated equilibrium*.[14]

The 'reactions' or 'dispositions to judgement' that act as inputs for the game should be something more than just 'brute' 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 of 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, depending on the context.

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

[14] 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.

they give us the right dispositions, our solution is both *ad hoc* and perhaps also circular.[15]

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 second judgement. This follows from the very way we assume S learnt to use colour words.[16]

Accordingly, I will stipulate that only S's dispositions to judgement count as determining an equilibrium of a constitutive game of a supergame (Γ) 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.[17]

If, on the other hand, S were to have the opposite set of dispositions and privilege his judgement under the electric light, where the tie appeared blue, over his judgement that it was green, we should say that S means something else by 'blue' and 'green' than we do. 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

[15] Thanks to Andrea Guardo for this point.

[16] This example is from Sellars (Sellars 1997, 37–39).

[17] 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.

dispositions to judge that $68+57=125$ are stable, as well as those concerning the length of the two lines.

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 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 $(\Gamma)_{+}$. Each such second-order equilibrium path thus represents *one* possible interpretation of the symbol being used in the basic constitutive practice.

The game-theoretic structure of the practice is then able to pick out one such second-order equilibrium path as being actual in the practice, since 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 $(\Gamma)_{+}$ that can be selected, namely the one that the agent's dispositions to judgement agree on, and hence only one second-order equilibrium path through $(\Gamma)_{+}$ itself. 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. There will be one equilibrium path through $(\Gamma_t)_1$, one through $(\Gamma_t)_2$ and so on, for each supergame in $(\Gamma)_+$, and hence, only one second-order equilibrium path through $(\Gamma)_+$.

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 $(\Gamma)_+$. 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.

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.

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 discussion 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.

The same point applies to all contentful mental states that S might be in, his beliefs, wishes, understanding, etc.: their correctness conditions and content are given by a basic constitutive practice. This also explains how S can misapply a word: S's intention, however that is specified, 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.[18] It follows that if S did mean *plus* by his utterance of 'plus', then S *should* say '125' when asked what "57 + 68" is *in order* to be in accord with his own intentions at the time of utterance, since S's intentions are given content by the practice. Otherwise S would simply not have acted in

[18] 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".

accordance with their own intentions.[19]

What then about guidance by meaning? On this account, basic constitutive practices are not rule-governed, and so any constitutive rule that an agent might appeal to in explaining his actions, for example,

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

would derive its meaning from the basic constitutive practice S is embedded in, and not the other way around. The rule 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. 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.

Above I stated that I was hesitant to claim that this account of meaning is conventionalist one, or perhaps rather that I should be hesitant to say that constitutive practices are conventions. There are several reasons for this. The first is that on most definitions of *convention* (Marmor 2009; Vanderschraaf 2018; 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. Since I'm analysing meaning as *constitutive practices*, it is unclear what this means. If our practice of using a particular symbol was different, then its meaning would change, but it would thereby not be the same practice and therefore not the same meaning. This notion of arbitrariness is therefore not entirely clear in this case.

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

[19] The account is thus an instance of social externalism about content. The *loci classici* are Burge 1979, 1986.

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.

2.4 A worry about dispositions

On the present account, the fact that determines the meaning of S's utterances is a function of two factors, the dispositions to judgement of all the agents taking part in a constitutive practice and the game-theoretic structure of that practice. 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 (see e.g. Kripke 1982, 27).

In our case, I think we can explain how agents can have the necessary dispositions without making too heavy psychological demands on the agents. 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.

We might then say, following Jared Warren (Warren 2020, 9), 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 is a structure to the technique and so even if S does not have infinitely many dispositions, 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 $n + 1$.[20] A snapshot of how one such particular technique looks is something like the following calculation:

[20] In a way, one could read Warren's account as a reply to precisely this objection: how can S's dispositions be structured to cover enough cases? I do no think, however, that Warren's account succeeds in giving the correctness conditions of the use of concepts, which is what has concerned me here. See below for further discussion.

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

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 $n + 1$.

The question then becomes twofold: how can S have enough dispositions to carry to cover enough cases and what ensures that S is performing the right calculation, the one for addition and not for quaddition? 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?

The answer to those questions is 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 and say that whatever *that* is, 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 the case of *our* practice, *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. 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 demand does not place unreasonable psychological demands on the agent.

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 $n + 1$. 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] And that is all that we require to be able to account for the basic constitutive practice of using the word ‘red’ which gives its meaning.

2.5 Objections to communitarian accounts

What then about worries that have been raised against communitarian solutions to the paradox, including the sceptical solution? In particular, the worry that the communitarian solution results in an account where we cannot speak of the community going wrong? Surely, a satisfactory account should make some room for the notion of the community as a whole making a mistake? The following is Anandi Hattiangadi’s criticism of the sceptical solution, which is fairly typical in this regard and generalises to most communitarian solutions:

[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.

> 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, 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. By appealing to the dispositions of 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.

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, 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. More importantly, 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 no question about the community's uses tracking such a relation from past used to novel cases at all, since one second-order equilibrium path is selected immediately.

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. The same would happen if the make-up of the community changed. The account does therefore not require that the community itself is stable.

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 to which concepts our words refer, and in this sense we might say that meaning is an investigation-dependent property. But why shouldn't the meaning of *our* words depend on us? It is however not the case that properties *in general* are investigation-dependent on this account: it is up to us what the meaning of the term 'red' means 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 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'.[22]

[22] This is what I believe Wittgenstein means when he writes (PI, §§241–242):

> "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.

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, 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.

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.

If, however, visual impression is the *only* thing they go by, there isn't 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

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.

wrong.[23] 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.[24] 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. It is not 'up 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. We 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).

2.6 Does the account assume too much?

One might think that the account makes too substantial assumptions in order to work, assumptions that render it circular or question-begging. There are a few distinct worries here. The first such worry I want to consider is the possible objection that the game-theoretic structure we are using assumes a notion of rationality that requires the agents to already have the necessary concepts to be able to reason—that for an agent to know what option to pick, they already need to have *some* concepts, and if the account is supposed to explain the constitution and acquisition of concepts, it is circular.

This worry is natural, given how game-theory is often presented, but in

[23] Thanks to Carrie Jenkins and Crispin Wright for the objection this example raises. I would employ a similar strategy for Boghossian's "horsey cow" case, which is structurally similar (Boghossian 1989, p. 535–536).

[24] 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.

fact, it does not follow. Game-theory has been used, for example, to explain certain phenomena in evolutionary biology, e.g. how bacteria interact and evolve.[25] There, the pay-offs are not defined relative to what the agents themselves, i.e. the bacteria, subjectively consider good for themselves and the strategies are not chosen after rational deliberation, after all, bacteria are not even conscious, but instead, the formalism assigns a pay-off relative to the reproductive fitness that a given strategy affords the bacteria.

In our case, the agents are not guessing or deducing what the equilibrium will be, from their perspective there is always only one possible option: the one that fits with their stable judgement about a given case. The action that they take is in basic cases is only determined by their dispositions to judgement, and not based on any reasoning or rational decision-making.[26] As such, the practice itself is opaque to the agents taking part in it.

This is possible, because while an equilibrium path is defined by the actual actions taken by the agents (or in our case, actual dispositions to judgement), the actual equilibrium isn't. The agents are following a strategy of just doing what they are disposed to do, and *that* strategy system is in equilibrium, again, because of our assumption that the agents are similar enough to form similar dispositions after training. This also answers a similar worry, that the account assumes that the agents follow rules when acting, and hence that we assume rule-following when explaining rule-following. That is not the case, as we only require that they do what they are disposed to do when taking part in a basic constitutive practice, not that such practices themselves are rule-governed. Rules and meaning presuppose basic constitutive practices on this account, not the other way around.

This leads to a another worry: The account just outlined assumes that the agents taking part in a given basic constitutive practice form dispositions to judgement about future cases after training and instruction, and that this formation of dispositions is in some sense uniform for all the agents in the

[25] See Lambert, Vyawahare, and Austin 2014 for an example.

[26] Compare the following two remarks by Wittgenstein:

> The origin and the primitive form of the language game is a reaction; only from this can more complicated forms develop. (*Culture and Value*, p. 31e)

and

> I want to regard man here as an animal; as a primitive being to which one grants instinct but not ratiocation. Language did not emerge from some kind of ratiocation. (*On Certainty*, §475)

practice. One might then wonder if the account, thus described, does not already presuppose the solution—is there anything more needed than such primitive tendencies to form dispositions that extrapolate in uniform ways to new cases? And furthermore, isn't this a move that Kripke has already argued against with his arguments against dispositionalist accounts?

Let's first answer the second question. In the beginning of the paper, I outlined the problem in terms of a subject learning a language by seeing a finite number of exemplars, in order to draw attention to certain features of the problem. Now suppose S is learning a new concept, one that we, the philosophers in the metalanguage reasoning about S, do not know either. Suppose that this concept is a colour concept. S has seen a number of samples of this colour and now says to himself: "I shall call this colour 'bleikr'".[27] Now, S has some dispositions to judgement about which things count as 'bleikr', but what is the criterion of correctness here? When has S correctly judged that an object is bleikr?

Theorists have often assumed that the problem can be solved by finessing S's dispositions in such a way that they match a given concept—so that S means *red* by 'red' if and only if S is disposed to apply 'red' to red things (for a recent account of this nature, see Warren 2020). But as this simple example shows, this assumes that we, the philosopher's reasoning about S's dispositions, already have a grip on which concept 'red' refers to in the metalanguage—essentially that the paradox is already solved in the metalanguage. By looking at the problem from the point of view of language learning, however, we can see that this answer is inadequate, since presumably we also want to solve the paradox for our own language, with no metalanguage to appeal to. We cannot assume, when solving the problem, that there already exists a correct sameness relation from exemplars to new cases that our dispositions merely have to track. The real problem is about how such a relation is picked out in the first place, and mere dispositions

[27] 'Bleikr' is just an arbitrarily chosen word, meant to have no prior connotations for the reader.

cannot do this.[28]

Hence, if *S*'s dispositions are the only criterion for correctness of the application of the term 'bleikr', then it would follow that no matter what dispositions *S* has, *S* has the right dispositions to judge that something is indeed bleikr, and hence that *S*'s dispositions are by definition the correct ones, no matter what they are. The current account solves this problem by placing *S* in a basic constitutive practice of using the term 'bleikr' which *does* pick out one such relation from the past exemplars to future cases as correct, namely the one that lies on the second order equilibrium of the practice. Here, there is therefore the possibility of *S* not acquiring the right dispositions to judgement, namely if the do not match that of the equilibrium.

This also shows why we need more than one agent in the practice and why the game-theoretic structure of the practice is necessary—a trivial practice with just one agent would collapse into a pure dispositionalist account, where there would be no standard of correctness,[29] and similarly, if we just went by the dispositions of the agents, without the game-theoretic structure, the account would be left vulnerable to the kinds of objections community accounts have been subjected to in general.

What about the first question, then? Can we really assume that the agents form uniform dispositions to judgement without begging the question? First of all, we can safely assume that agents form *some* dispositions to judgement after training and instruction. This does not assume anything more than our agents reacting to some stimuli when they are present, like a neural network that has been trained on some dataset or an animal that has been trained to do tricks. This part of the assumption does not assume more than what pure dispositionalist accounts already assume. Of course, my claim is not that this is what learning a language consists in, merely that

[28] 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. It is rather that if we move the problem onto ourselves, instead of a different, hypothetical subject like *S*, we can more easily see why dispositions are unable to set up a standard of correct use, no matter how well they track a concept.

[29] There is therefore an implicit argument against 'private language' in the account: if the terms of *S*'s language were only in principle understandable by *S*, then *S* could not be taking part in the basic constitutive practices of using the terms, and hence there would be no correctness or incorrectness in S's language.

our assumption does not require more than this.

What about the our reliance on the agents forming *uniform* dispositions—is that not an illegitimate use of 'uniform' which begs the question? Doesn't our use of 'uniform' already presuppose that the problem is solved, since we cannot say that the agents get a positive pay-off unless they have the same dispositions, and therefore we have a notion of 'same', the very problem we aim to solve? First of all, consider the meeting game used as an example above. There, we stipulated that the agents get a pay-off when they meet at the same restaurant. This notion of 'same' however doesn't depend on any prior notion of sameness that the agents have, nor us, the philosophers reasoning about them in the metalanguage. The reason is simply that if the agents are not physically in the same place, then they will not get the pay-off of having dinner with each other. This notion does not depend on any prior description of what it is to go to the same restaurant. If we had stipulated, for example, that 'going to the same restaurant' meant 'if R goes to restaurant n, then C goes to some restaurant m such that $m \neq n$', then they simply would not have gotten they pay-off from enjoying dinner with each other, despite our stipulation. They are not doing the *same* in the sense that their—or our—use of concepts coincide, but in a way that depends on how the external world is and their aims in it.

We can view the account just presented as an attempt to spell out a distinction Wittgenstein speaks of often: that between agreement of opinions, on the one hand, and an agreement in action or judgements, on the other.[30] With that in mind, our assumption of uniformity is therefore one of 'being wired in the same way' in this primitive, unconceptualised sense: of just mechanically reacting in the 'same' way, whatever that is. It assumes a common human nature, in a biological sense, but also that the interactions between the agents, including each agent's induction into the practice, take place in an external world, independent of the agents.[31]

One might still ask, is this use of the metalanguage not illegitimate in the

[30] See for instance PI, §§241–242 or *Lectures on the Foundations of Mathematics*, pp. 183–184: "And it has often been put in the form of an assertion that the truths of logic are determined by a consensus of opinions. Is this what I am 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." (Wittgenstein 1976, p. 183-184).

[31] For discussion of this point, see Verheggen 2003, section 9.

same way as I argued that the way dispositionalists use the metalanguage is? No, because the description of that 'sameness' in the metalanguage doesn't actually matter, just that the agents are able to function in the practice—we register the pay-off in the metalanguage, if they are able to communicate, give orders and obey them, have beliefs and so on. My claim is not that *any* use of the metalanguage in describing the practice of the agents is illegitimate. If that were the case, then any possible solution would be circular in this way.[32] This assumption of a 'common form of life' is of course a substantial one, but crucially, it is not circular, nor does it assume that the problem is already solved.[33] Likewise, this assumption alone is not enough to solve the problem, since on its own it would be equivalent to dispositionalist accounts, especially in their community forms, and hence have all their problems.

3 A genuine case of rule-following

In what I've said above, it might 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?

[32] We could imagine, for instance, that the agents form dispositions in some regular way which is totally unlike ours, and that we would be unable to come up with any coherent description of their practice because we would never call what they do 'the same'. This situation wouldn't entail that they are not participants in basic constitutive practices.

The sceptic might still argue, however, that even if A and B both use what we call the same symbols in their practice, e.g. '125' in that given case, A might still mean *quus* by her use of that symbol, while B means *plus*. After all, one can mean different things by the same word.

However, if A and B both use the same symbol in every possible case and get on in their practice, can communicate, etc., then what does it mean that A *means* something else than B? A's 'meaning' seems to have dropped out as irrelevant.

[33] In fact, dispositonalists themselves must hold to some form of it, since otherwise their account could not explain the possibility of communication. If meaning is reducible to dispositions, then agents must, in some sense, have similar dispositions in order to speak to each other. This account does not assume similarity in a stronger sense than this. This point might even apply to primitivists about meaning.

Consider the two following cases of ostensible rule-following on what Crispin Wright has called 'the modus ponens model' of rule-following—a natural way of conceiving of rule-following in general. First 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:

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

(Premise) $\ldots x \ldots$

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 '$\ldots x \ldots$'. 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.[34] 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" (Wright 2007, 496) and that in such cases "we do not really *follow* – are not really guided by – anything" (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.[35]

The basic dilemma is, 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

[34] And of course, if we were to put anything else than *red* in for '$\ldots x \ldots$' we could just take *that* concept to be basic and repeat the reasoning until the bottom.

[35] See also Wright 2012 for discussion.

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, nor in our explanation of the meaning of the symbol in question. 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.[36]

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 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—although of course much more needs to be said to give a full account of rule-following, a task which falls outside the scope of this paper.

[36] Glüer and 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.

However, the success of the present account to avoid the paradox presented by the modus ponens model gives us a strong reason to suspect that language is in fact not rule-governed. This, it should be noted, also explains why Searle's problem is not a problem for this account, since we should expect that every rule, both regulative and constitutive, has an associated basic constitutive practice or at the very least bottoms out in such a practice. In other words: both regulative and constitutive rules are given content by basic constitutive practices, off which constitutive rules can be trivially read, and as such are at the most basic level quite the same. The only difference is how they are described. However, this is not a point that applies to the basic constitutive practices themselves, since the account does not rely on any such distinction.

4 Concluding remarks

In this paper, I've offered an account of rule-following in terms of what I've called basic constitutive practices. On this account, the correctness conditions for applying a term to an object 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. Thereby, what it *is* to be F-ing is determined by the equilibrium and therefore also what it is to correctly use the term F, and since this determination only requires that the agents form dispositions through their linguistic training and that their practice of using a term has a certain structure, a sameness relation from past exemplars to future cases is picked out without circularity, thus providing an answer to the paradox.

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.

Finally, I've argued that this account gives us a strong reason to suspect that language isn't in fact rule-governed at all. Meanings explain rules, not the other way around, and by appealing to basic constitutive practices, we can nevertheless account for the normative character of meaning—its correctness conditions—without a further appeal to rules.

Acknowledgements

I would like to thank Peter Sullivan and Crispin Wright for their insightful comments and extensive discussions on the content of this paper. I would also like to thank audiences in Stirling, St Andrews and Milan for their comments, as well as the anonymous reviewers who read the paper before publication. It has been much improved by their input.

Funding

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

References

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.

Bloor, David. 1997. *Wittgenstein, Rules and Institutions.* London: Routledge.

Boghossian, Paul. 1989. "The Rule-Following Considerations." *Mind* 98, no. 392 (October): 507–549.

Burge, Tyler. 1979. "Individualism and the Mental." *Midwest Studies in Philosophy* 4 (1): 73–122.

———. 1986. "Individualism and Psychology." *The Philosophical Review* 95 (January): 3–45.

Gintis, Herbert. 2009. *The Bounds of Reason.* Game Theory and the Unification of the Behavioural Sciences. Princeton: Princeton University Press.

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. Palgrave-Macmillan.

Guala, Francesco, and Frank Hindriks. 2014. "A Unified Social Ontology." *The Philosophical Quarterly* 65 (259): 177–201.

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.

Kripke, Saul. 1982. *Wittgenstein on Rules and Private Language.* Harvard University Press.

Kusch, Martin. 2002. *Knowledge by Agreement: The Programme of Communitarian Epistemology.* Oxford: Oxford University Press.

Lambert, Guillaume, Saurabh Vyawahare, and Robert Austin. 2014. "Bacteria and game theory: the rise and fall of cooperation in spatially heterogeneous environments." *Interface Focus* 4 (20140029).

Lewis, David. 1969. *Convention: A Philosophical Study.* Wiley-Blackwell.

Marmor, Andrei. 2009. *Social Conventions: From Language to Law.* Princeton: Princeton University Press.

McDowell, John. 1984. "Wittgenstein on Following a Rule." *Synthese* 58 (3): 325–363.

Millar, Alan. 2002. "The Normativity of Meaning." *Royal Institute of Philosophy Supplement* 51:57–73.

Peregrin, Jaroslav. 2012. "Inferentialism and the Normativity of Meaning." *Philosophia* 40 (1): 75–97.

Rawls, John. 1955. "Two Concepts of Rules." *The Philosophical Review* 64 (1): 3–32.

Reiland, Indrek. Forthcoming. "Constitutive Rules: Games, Language, and Assertion." *Philosophy and Phenomenological Research.*

Ruben, David-Hillel. 1997. "John Searle's The Construction of Social Reality." *Philosophy and Phenomenological Research* 57 (2): 443–447.

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.

Sillari, Giacomo. 2013. "Rule-Following as Coordination: A Game-Theoretic Approach." *Synthese* 190 (5): 871–890.

Vanderschraaf, Peter. 1995. "Convention as Correlated Equilibrium." *Erkenntnis* 42 (1): 65–87.

———. 1998. "Knowledge, Equilibrium and Convention." *Erkenntnis* 49 (3): 337–369.

Vanderschraaf, Peter. 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.

Warren, Jared. 2020. “Killing Kripkenstein’s Monster.” *Noûs* 54 (2): 257–289.

Whiting, Daniel. 2007. “The Normativity of Meaning Defended.” *Analysis* 67 (2): 133–140.

———. 2009. “Is Meaning Fraught with Ought?” *Pacific Philosophical Quarterly* 90 (4): 535–555.

———. 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.

Wittgenstein, Ludwig. 1976. *Wittgenstein’s Lectures on the Foundations of Mathematics.* Edited by Cora Diamond. Chicago: University of Chicago Press.

Wright, Crispin. 2001. “Does *Philosophical Investigations* §§258–60 Suggest a Cogent Argument against Private Language?” In *Rails to Infinity: Essays on Themes from Wittgenstein’s Philosophical Investigations.* Cambridge: Harvard University Press.

———. 2007. “Rule-Following Without Reasons: Wittgenstein’s Quietism and the Constitutive Question.” *Ratio* 20, no. 4 (December): 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 University Press.