Rule-following as Coordination: A Game-theoretic Approach

Giacomo Sillari, Philadelphia, Pennsylvania, USA

> Make the following experiment: *say* "It's cold here" and *mean* "It's warm here". Can you do it?
>
> Ludwig Wittgenstein, *Philosophical Investigations*, §510.

> I can't say "it's cold here" and mean "it's warm here"—at least, not without a little help from my friends.
>
> David Lewis, *Convention*.

1. Rule-following, coordination and normativity

The slogan that "meaning is normative" is best understood in the context of strategic interaction in a community of individuals. Famously, Kripke has argued in (Kripke 1982) that the central portion of the *Philosophical Investigations* describes both a skeptical paradox and its skeptical solution. Solving the paradox involves the element of the *community*, which determines conditions of assertability in the language. A battery of argument is used to show that meaning (or, in general, rule-following) cannot be explained by resorting to an individual's mental states, or her past use, or her dispositions. By exclusion, this indicates that no descriptive fact is constitutive of meaning, and hence that "meaning is normative." Arguably, the normativity of meaning stems from the assertability conditions holding in the society (indeed, membership in the community depends on one's record of compliance.) But *how* exactly is the existence of such conditions sustained in the community? And is it accurate to say that there is no *fact* to the matter of rule-following?

I need an important *caveat* here: To answer these questions, I momentarily step back from analyzing *meaning* and elaborate on the more general notion of *rule-following* instead. Kripke himself uses the terms meaning and rule-following rather interchangeably in (Kripke 1982). I will conform to the ambiguous usage for ease of exposition, and mention my justification for it in the last section of this contribution.

Wittgenstein states (§§198, 199) that a rule is followed *insofar* as there exists a custom, a convention. I argue that this and similar remarks in the *Philosophical Investigations* are illuminated when looked at through the lens of David Lewis's theory of convention. Lewis argues in (Lewis 1969) that coordination games (situations of strategic interaction in which the interest of the players roughly coincide) underlie every instance of convention, in that a convention is a regularity in the solution (equilibrium) of recurrent coordination games. The agents participating in the convention conform to the regularity because they prefer conformity over non-conformity, conditional on other agents' conforming. They form the belief about other agents' conformity through some coordination device: explicitly—through agreement—or tacitly—because a certain action stands out as the one that most likely (almost) everyone will pick. Such an action is *salient* to the parties. In the case of a recurrent coordination problem, a special kind of salience—*precedent*—is at play.

Conventionality in the sense of Lewis is sufficient for some degree of normativity to arise. Indeed, in a community in which a certain custom is in place—say, the custom of going by sign-posts—there is an equilibrium in the actions and beliefs of the agents involved such that the agents prefer conformity to the custom, provided that all other members in the community act according to the convention. If I do *not* go by sign-posts, or I go by them in a funny, abnormal way (for instance, going in the direction opposite to the one indicated) I act contrary to both my preferences—because I will not get where I intend to go—and the preferences of other members of the community—because, say, I will end up being late, or not showing up at all. My reputation will suffer. This indicates that, in general, parties to a convention feel, to a larger or smaller extent, the pressure to conform. As Lewis puts it, conventions are a kind of social norm. But are we entitled to cast the rule-following phenomenon in a game-theoretic account of convention?

In its most general terms, the communitarian view maintains that, while many interpretations of a given rule may arise, there is (roughly speaking) only one correct way to abide by the rule, as determined by the community. In particular, the customary action is the action that accurately corresponds to the rule. The problem with arguments of this general form is that the *same* skeptical paradox meant to show the impossibility of solipsistic rule-following applies to the community. *Which* is the customary action? And *why*? Past use is no sufficient grounds to answer such questions for the community, as it is not sufficient grounds in the solipsistic case, since the community can come up with a variety of interpretations of the rule, just as well as the individual can. However, if we introduce a *strategic* element in the behavior of community members, then the skeptical paradox disappears (or, as we shall see in the next section, gets "pushed towards bedrock.") If we interpret rule-following as *coordination equilibrium* in a coordination problem, then there *is* a clear and compelling fact to the matter of what "going by the rule" consists of. In particular, individuals (and the population they interact with) who go by the rule net a higher payoff than do individuals who transgress the rule. Moreover, transgressing the rule comes at a price, both for the transgressor and for the agents interacting with him. Non-conformative behavior will end up being sanctioned (eventually with expulsion from the community), while conformative behavior will perpetuate itself, being based on the agreement to act according to given rules. In this sense, agreement is the agreement in preferences and beliefs that support a specific equilibrium in the recurrent coordination game.

Thus, the "little help" needed by Lewis from his friends in the answer to the challenge of §510 reported in the epigraph consists then in their *agreeing* to change their preferences and beliefs, switching in so doing from one solution of a recurrent coordination game to another. Consider §224:

> The word "agreement" and the word "rule" are *related* to one another, they are cousins. If I teach anyone the use of the one word, he learns the use of the other with it.

I believe that the view expressed in this section captures the sense in which "agreement" and "rule" are related: A custom—and hence a rule—does not hold without an agreement in preferences and beliefs—and hence in co-

ordinative, conventional actions—on part of the members of the community.

2. Precedent and justification

Although my interpretation of §224 surely appears contentious to many, it should become clear by the end of this section that in fact it jibes with the traditional reading. I find in §241 the cue to the traditional interpretation of "agreement" in §224:

> [Human beings] agree in the *language* they use. That is not agreement in opinions but in form of life.

Lebensform is a rich and profound philosophical concept that does *not* reduce to the preferences and beliefs (to the *opinions*) held in a community. Still, I claim that the notion of *Lebensform* is related to the Lewisian picture of conventional behavior and that preferences and beliefs in the community in fact spring from it.

Precedent lies at bedrock, where the spade is turned (§ 217) and one acts blindly (§ 219) conforming to the convention and obeying the rule. Without reliance on precedent, no conventional strategic interaction in the sense of Lewis is possible and, as I have argued in the previous section, without strategic interaction the community is in no better position than the individual is in determining which course of action is in accord with the rule. Indeed, as Margaret Gilbert tersely points out in (Gilbert 1990), in Lewis's account of convention practical rationality does not yield any justification to act in conformity to precedent. She argues that, on the contrary, conformative action is *blind* in the Wittgensteinean sense. Consider the two person case: Given their conditional preference, one is justified in conforming if she believes that the other will conform. But the other will be justified in conforming if he believes that the first individual will. Thus, she will be justified if she believes that he believes that she will, and so on. In the endless replication of each other's reasoning, at no point anyone will come to have sufficient reason to conform.

I have argued elsewhere (Sillari 2005, 2008) that in fact precedent gives rise to the series of replications about hypothetical future conformity, which, in turn, *inductively* ground for both individuals the first-order belief that the other will conform. There is no deductive, infallible passage from past to future conformity. There rather is a causal, inductive one (cf. the interlocutor in §198: "[…] What sort of connexion is there [between the expression of a rule and my actions]?—Well perhaps this one: I have been trained to react to this sign in a particular way, and now I do so react to it.[—]But that is only to give a causal connexion […]") Wittgenstein speaks of "blind action", Gilbert speaks of an "a-rational tendency". For (McDowell 1984), understanding is "precarious and contingent", in that there is no guarantee that my grasping a concept will continue working tomorrow as well. No strong, logical, deductive nexus is to be found between precedent and future conformity. Rather, the relation between precedent and future conformity lies at bedrock, as pointed out in §481:

> If anyone said that information about the past could not convince him that something would happen in the future, I should not understand him. […] If *these* are not grounds, then what are grounds?

As flimsy as the relation might be, we all endorse it since, as the traditional interpretation of §224 indicates, we all share an agreement in *Lebensform*. Our systems of concordant beliefs about each other conformity *stem* (albeit not deductively) from such a fundamental agreement. In turn, from our concordant beliefs and conditional preferences stem our conventions and customs, and hence our capacity to obey or to go against a rule.

The game-theoretic analysis of rule-following reveals that preferences and beliefs of community members strategically determine what course of action is in accord with the rule. The formation of beliefs, however, is a bedrock notion. Can a game-theoretic analysis help us reduce the phenomenon of rule-following any further? It is well known that Wittgenstein invites us not to dig under bedrock. To ask whether it is possible, and how it may be done, I finally tackle the issue of the relation between meaning and rule-following and turn to the final section of this contribution.

3. Meaning and rule-following

In this section I focus on *meaning* by looking at a special case of coordination problems involving communication. Rather than attacking the question of meaning in *language* (a question that lies well beyond the scope of this note) I will look at the simpler case of meaning in *signaling systems* (cf. Lewis 1969). Signaling systems are a special case of coordination problems. In a signaling game certain actions (performed by the *audience*) correspond to certain states of the world (observed by the *speaker*.) The speaker sends a *signal* depending on what state of the world she observes. The audience performs a certain action depending on what signal she receives. Both speaker and audience prefer that the action corresponding to the actual state of the world be performed. For that to happen, they need to coordinate their strategies (which for the speaker are functions linking states to signals for the speaker, while they are functions linking signals to actions for the audience.) When coordination is achieved, then, a signal may assume the indicative meaning that "the state of the world is such-and-such" or the imperative meaning "perform such-and-such action!" depending on further characteristics of the situation that need not concern us here. The relevant point is that signaling problems are a special kind of coordination problems.

The builder-assistant language-game of §2 is a clear example of a signaling game that one surmises Lewis might have had it in mind when characterizing the class of signaling games: The builder is the speaker. She observes, for instance, the state of the world in which she needs a slab and she sends the signal "Slab!". The assistant is the audience. He receives, in this example, the signal "Slab!" and performs the action of bringing a slab. The *caveat* issued in the opening section of this paper can now be lifted. If rule-following is conventional action, then meaning, as the by-product of conventional action (signaling) is a special case of rule-following.

In the case of linguistic coordination, the skeptical paradox can be understood as an instance of the problem of indeterminacy of meaning. David Lewis, in *Convention* (cf. pp.199-200) as well as in later works, has tackled the problem. In particular, in (Lewis 1992) he confronts "Kripkenstein's challenge (formerly Goodman's challenge)" (p. 109) and argues that for sentences never uttered before, the rules governing the used fragment of the language determine the rules for the unused portion, too. The argument is that they do so because although extrapolation from used fragment to unused portion is "radically underdetermined", only a *minority* of extrapolations are *straight*—and acceptable—while the

vast majority are *bent*: gruesome, gerrymandered—and disposable. The argument carries over also to the extrapolations we all perform daily from precedent to current use. “Straightness” of extrapolation, or of grammar, is a bedrock notion, on which we all agree. Lewis warns us that digging under bedrock—analyzing straightness of extrapolation—cannot be a linguistic enterprise, since our use of language depends on it in the first place. Digging under bedrock points, therefore, to the *ontological* distinction between properties that are natural and properties that are not.

Literature

Wittgenstein, Ludwig Philosophical Investigations, 1953

Kripke, Saul Wittgenstein on Rules and Private Language, 1982

Lewis, David Convention: A Philosophical Study, 1969

Lewis, David “Meaning without use”, Australasian Journal of Philosophy, 1992

Gilbert, Margaret “Rationality and Salience”, Philosophical Studies, 1990

McDowell, John “Wittgenstein on Following a Rule”, Synthèse, 1984

Sillari, Giacomo “A Logical Framework for Convention”, Synthèse, 2005

Sillari, Giacomo “Common Knowledge and Convention”, Topoi, 2008