UNDERSTANDING INSTITUTIONS: THE SCIENCE AND PHILOSOPHY OF LIVING TOGETHER

FRANCESCO GUALA

Reviewed by Christopher Clarke

Understanding Institutions: The Science and Philosophy of Living Together
Francesco Guala
Princeton: Princeton University Press, 2016, £27.95
ISBN 9780691171784

If you take a philosophical interest in social ontology, either as a specialist or as a non-specialist, then this book is essential reading.

In Chapters 1—4 and 7, Guala clarifies and defends his favoured approach to social institutions. The concepts at the centre of his approach are drawn from game theory: individual preferences and motivations, individuals' expectations, and game-theoretic equilibrium. Students and non-experts will benefit from Guala's careful and clear exposition. And they will also welcome his sustained use throughout the book of three stylized examples: the institutions of property, of money, and of marriage. Experts, in contrast, will already be very familiar with game-theoretic approaches such as Guala's. Nevertheless, many will value the sharpness with which Guala characterizes his account, carefully spelling out how concepts such as 'rule' can be rigorously defined and analysed using more basic game-theoretic concepts.

In Chapters 5, 6, and 8, Guala clarifies his account of institutions by contrast with others, such as those of John Searle and Margaret Gilbert. In contrast to game-theoretic accounts, the central concepts here are constitutive

rules, normativity, and collective intentions. Guala's clarificatory work in this section is where the book really advances our understanding. Guala uses some precise definitions and illuminating examples to teach us exactly where the prevailing accounts of institutions agree and disagree. No easy task. What's more, Guala develops a very powerful argument against Searle's theory of constitutive rules, an argument that deserves to be taken very seriously. This part of the book offers the non-expert an accessible overview of the extensive theoretical literature on cooperation.

In Chapters 9 to 14, Guala teases apart the various ways in which institutional kinds may or may not be socially constructed. This clarificatory work allows Guala to defend a realist theory of institutional kinds. But if realism about institutional kinds is true, Guala argues, then this has important consequences: the relationship between social ontology and social reform is very different to that supposed by some philosophers. Guala's chief targets are those philosophers who follow Sally Haslanger in doing social ontology in an 'ameliorative' vein. Much of Guala's expository work here is very helpful, carefully drawing connections between the literature on natural and social kinds on the one hand, and on institutions and cooperation on the other. On the whole, this book makes a substantial and welcome contribution to the literature on social ontology. I will turn now to some more detailed (and more critical) remarks.

A Renault and a Citroen are rapidly approaching a junction in the streets of Paris. The two drivers face a 'problem of coordination' in virtue of the following two facts: First, for the Renault driver, the action she prefers to take at the junction depends on what the Citroen driver will ultimately do at the junction. If the Citroen driver continues through the junction, the Renault driver would prefer to stop; but if the Citroen driver stops, the Renault driver would prefer to continue. And indeed the converse point holds for the Citroen driver's preferences. Second, there is some alignment in the preferences of the two drivers: neither driver wants to crash, and so the least preferred outcome for each driver is that both cars continue through the junction. That's not to say that the two drivers' preferences align perfectly; each driver prefers that the other yields.

One 'solution' to this problem of coordination is that the drivers adhere to the following regular pattern of behaviour: if the traffic light facing you is RED, then stop; if it is GREEN, then go. This pattern of behaviour counts as a solution to the problem of coordination because of two facts. First, both drivers prefer this pattern of behaviour to some alternative pattern of behaviour (for example, always go, no matter what). Second, this pattern of behaviour is a Nash equilibrium: assuming that the Citroen driver will adhere to this pattern, the Renault driver prefers to adhere to this pattern too; and assuming that the Renault driver will adhere to this pattern, the Citroen driver prefers to adhere to it too. More specifically, this pattern of behaviour is a 'correlated equilibrium', in that the pattern of behaviour involves individuals making their behaviour dependent on a public signal, namely, the traffic light.

Another important fact to note is that the adoption of this solution to the coordination problem is contingent. Had things been different, the pattern of behaviour exhibited at this junction in Paris would have been different. For example, the pattern might have been: if the traffic light is BLUE, then stop; if it is PINK, then go. Or it might have been: if the other car is larger, then stop; if it is smaller, then go. And, in part, this contingency is due to our remarkable flexibility as humans: we can mentally represent a huge variety of potential patterns of behaviour, and we can allow our own behaviour to be guided by these representations.

I offer this traffic light example as a quick illustration of Guala's definition of institutions. An institution is a regular pattern of behaviour that (i) is a correlated equilbrium solution to a coordination problem, (ii) is relatively stable over time, but nevertheless (iii) is contingent, and (iv) is mentally represented by the participants.

Guala places great emphasis on the fact that his definition of institutions employs the notion of a correlated equilibrium, whereas standard game-theoretic definitions employ the notion of a Nash equilibrium. This is a distinction without much of a difference, I think. Whether a solution counts as a Nash equilibrium or as a correlated equilibrium is usually just an artefact of how one chooses to formally model the situation. As Guala himself notes, all correlated equilibria can be satisfactorily modelled as Nash equilibria by an easy modification to one's formal game-theoretic characterization of the situation. Thus it's actually conditions (iii) and (iv) that distinguish Guala's definition of institutions from other game-theoretic definitions.

With institutions defined in this way in Chapters 1 to 4, the task of Chapter 7 is to survey the literature on the stability of institutions. This literature seeks a more complete explanation of why institutions such as money, marriage, and private property have been so stable over time. Why might theorists want a more complete explanation of this? Well, a behavioural regularity is stable when the individuals adhering to the regularity continue to be motivated to adhere to the regularity. But, for a regularity that is an equilibrium solution to a coordination problem, individuals are motivated to adhere to the regularity insofar as they expect the other individuals involved to adhere to the regularity. Therefore, to fully understand the stability of institutions one needs to understand the social and psychological mechanisms whereby individuals form stable expectations about other individuals' behaviour.

One approach, following David Lewis, is to say that the Renault driver's (first-order) expectations about the Citroen driver's behaviour depend upon the Renault driver's (second-order) expectations about how the Citroen driver expects the Renault driver to behave; and these second-order expectations depend in turn upon third-order expectations, and so on. According to Lewis, this complicated system of expectations is stabilized by a 'public event', for example the public event of the French government announcing in the newspapers that traffic lights are being installed, and instructing Parisians that if the traffic light facing you is RED, then stop; if it is GREEN, then go.

Guala presents and clarifies a powerful challenge to Lewis's account. He then offers Adam Morton's account of expectation formation as an alternative, which does not appeal to higher-order expectations, and which is superficially similar to Bacharach's and Gold and Sugden's 'team reasoning' account of coordination. According to Morton, the Renault driver thinks that 'if the traffic light facing you is RED, then stop; if it is GREEN, then go' is the obvious solution to the coordination problem. The Renault driver then engages in simulation. She puts herself in the Citroen driver's shoes. In doing so, she implicitly assumes that the Citroen driver is culturally and psychologically like her in several respects, in particular, implicitly assuming that the Citroen driver also thinks that the above regularity is the obvious solution. As a result, the Renault driver expects the Citroen driver to stop if the light is red, and to go if the light is green.

I think Guala's assessment of the relative merits of Lewis's account versus Morton's is misleading. This is because Guala ignores one crucial shortcoming of Morton's account, namely, that Morton's account is incomplete. Morton does not tell us much about the circumstances under which an agent will feel that a particular solution to a coordination problem is the 'obvious' solution. As it stands, Morton's approach can merely gesture loosely to 'culture' or 'historical precedent' to explain why individuals take a solution to be obvious. But these appeals to historical precedent and the like are really just place-holders for a comprehensive social and psychological explanation of how the above expectations are formed (as Guala himself points out when criticizing some of Lewis's contributions to the literature). Thus Guala overstates the merits of Morton's account.

Other approaches to institutions place three concepts centre stage: constitutive rules, normativity, and collective intentionality. Thus each of these three concepts marks a point at which Guala's approach to institutions seems to be threatened by a rival approach. How does Guala respond to these threats?

With respect to the notion of collective intentionality, Chapter 8 argues that collective intentions ('we intend to move this table up the stairs' or 'I intend that we move this table up the stairs') are not necessary for the existence of institutions. Slavery is an institution, but enslaved humans do not usually form collective intentions with those who enslave them. Guala also points out some problems with the 'team reasoning' account of collective intentionality, problems not shared by Guala's preferred account of coordination, he claims.

With respect to Searle's notion of constitutive rules, Chapter 5 argues that the distinction between constitutive rules and regulative rules is a trivial distinction. Consider the following constitutive rule, for example: (1) 'An object counts as a traffic light in Paris, if the object has three coloured lights and it is beside a dashed line in the road'. Following Frank Hindriks, Guala argues that anyone who accepts a constitutive rule such as (1), also needs to accept some regulative rule such as: (2) 'If the traffic light facing you is RED, then stop; if it is GREEN, then go'. Why? Well, without accepting a regulative rule such as (2), the concept of a 'traffic light' would be meaningless, and thus the constitutive rule (1) would be meaningless too. It follows that the mental state X of accepting rule (1) is identical to the mental state Y of accepting rule (1) in combination with a rule like rule (2). But, Guala continues, this latter mental state Y is just the mental state Z of accepting the following rule: (3) 'If you are in Paris, and there is a three coloured light, and a dashed line on the road, then stop if the light is red, but go if the light is green'. By transitivity of identity, it follows that the mental state X of accepting a constitutive rule such as (1) is identical to the mental state Z of accepting a regulative rule such as (3). So there is no genuine distinction between constitutive rules and regulative rules. Thus, contrary to what Searle says, his theory of institutions is not a rival to the game-theoretic approach that Guala favours, at least not in this respect. This is an extremely powerful argument against Searle's theory.

Finally, with respect to normativity, there are two potential points of disagreement between game-theoretic approaches to institutions and alternative approaches. An illustration of the first point of disagreement is the following question: to what extent do individuals desire to adhere to the behavioural regularity 'tell the truth'? One 'hard-nosed' view is that any given individual desires to tell the truth only insofar as doing so serves her own personal interests (or the interests of her friends and family). For example, she might tell the truth in order to avoid punishment for lying, or in order to further some cooperative enterprise that serves her interests. The alternative to this hard-nosed view is that some (or many) individuals also desire to tell the truth as an end in itself. Now for point of disagreement number two: what is the relationship between desire and motivation? One 'Humean' flavoured view is that individuals are only motivated by their desires. Individuals are motivated to tell the truth only insofar as they desire to tell the truth (either as an end in itself, or as a means to some other end). The alternative to this Humean view is that an individual could be motivated to tell the truth, even absent any desire to tell the truth. For example, she might strongly believe that she ought to tell the truth or that she is morally obliged to tell the truth. And these beliefs might motivate her, independent of any of her desires.

The game-theoretic tradition has been associated more with the hard-nosed view of what people desire, and with the Humean view that ultimately only desires motivate people to action. What Chapter 6 does is point out that the game-theoretic approach to cooperation can be easily divorced from these two views. On Guala's interpretation of game theory, when game theorists talk about an agent's payoff or preferences, they are summarizing the agent's motivations, whether these motivations are just desires, or whether they also include motivational beliefs. Furthermore, there is no assumption that the agent desires only what is in her own interests (or the interests of her friends and family). Thus, Guala shows, an attack on the hard-nosed view of desire or an attack on the Humean theory of motivation are not threats to the game-theoretic approach to institutions. Showing this, I think, contributes a lot to understanding the proper relationship between game-theoretic accounts of cooperation and their rivals, such as Gilbert's and Searle's.

This achievement is obscured, however, by Guala's reliance on the phrase 'normative force'. In talking about 'normative force', he mixes together the ideas of (i) an agent being punished for deviating from a behavioural

regularity, (ii) an agent desiring to follow a regularity in and of itself, and (iii) an agent having some motivations that are independent of her desires. For this reason, I don't fully grasp Guala's way of drawing the distinction between norms and mere conventions: norms are conventions that are backed up by normative force (where 'normative force' can be represented game-theoretically using a 'delta parameter' in the payoff function).

To summarize, Guala's defence of the game-theoretic approach is both a daring attack (on the importance of collective intentionality and of constitutive rules) and a tactical retreat (on issues of normative force). It's for this reason that his account is, in my view, compelling and interesting. Guala says that the key feature of his account is that it unifies the account of institutions 'as rules' with the game-theoretic account of institutions as equilibria. But, in my view, this way of putting things undersells what Guala achieves.

One important thing that social scientists and social activists do is classify institutions: the government of the Netherlands and the government of Sweden, for example, are grouped together as belonging to the same type, labelled 'constitutional monarchies'. How should we group together and label institutions? Sally Haslanger's 'ameliorative' principle urges us to do this in the way that would best promote social justice. Take, for example, the social recognition in seventeenth-century Florence of many male-female partnerships and take also the social recognition in twenty-first-century Argentina of many partnerships between two partners of any gender. According to the ameliorative principle, we should group these two institutions together as belonging to the same type, and we should label this type 'the institution of marriage'. And we should do so because doing this would help fight the injustices faced by LGBT people.

Guala agrees that we should group together the twenty-first-century Argentinian institution and the seventeenth-century Florentine institution both as 'institutions of marriage'. But he disagrees with Hasslanger on why we should do this. Instead, he offers a 'realist' principle for classifying institutions that rivals Haslanger's ameliorative principle. We should classify institutions, Guala seems to say, in the same way that chemists classify chemical substances. First, institutions should be grouped together into objectively privileged kinds, just as chemical substances are grouped together based on their molecular components. And, second, we should use the traditional labels for institutions, Guala seems to say: take all the traditional exemplars of the label 'marriage' (such as the seventeenth-century Florentine institution) and identify the objectively privileged kind to which most of these exemplars belong; we ought to use the label 'marriage' to denote all and only the members of this objectively privileged kind. (I find it useful here to separate out Guala's approach into a realist principle for grouping institutions, and a realist principle for labelling them).

Why does Guala think that his realist approach entails that the LGBT-inclusive Argentinian institution should be grouped together with the LGBT-exclusionary Florentine one, under the label 'marriage'? Guala begins by saying that institutional kinds are functional kinds. All the institutions that are members of a given institutional kind are united by the fact that each institution has the same function: each institution offers a solution to the same type of coordination problem. But when do two coordination problems count as the same type of coordination problem? Think of the huge variety of institutions of higher education (each of which may or may not contain a medical school, or a business school, or a law school, or a philosophy department, or an anthropology department, or a department of public policy). Which of these institutions offer a solution to the same type of coordination problem? I'm not sure how to answer this question. And Guala doesn't say much to help us here. For the same reason, it's difficult to evaluate whether the Argentinian institution and the Florentine institution really do solve the same coordination problem, and so it's difficult to evaluate whether Guala's realist principles counts them as belonging to the same institutional kind, that is, 'marriage'. In short, I think Guala's treatment of marriage, although thought-provoking, is too vague to be compelling as it stands.

Putting this aside, I think Guala's approach faces a more troubling worry. Is there any sense in which democracy offers a solution to a different coordination problem than the one to which dictatorship offers a solution? If not, then Guala's realist principles have the absurd consequence that 'democracy' and 'dictatorship' pick out exactly the same institutions. Similarly, is there any sense in which money offers a solution to a different coordination problem than the coordination problem to which bartering offers a solution? If not, then Guala's realist principles have the absurd conclusion that 'money economies' and 'barter economies' pick out the same institution. At points, I get the impression that Guala would be inclined to respond to such worries by jettisoning his realist principles about how institutions should be grouped and labelled. Guala might say that what he ultimately wants to defend is the metaphysical thesis that there are (objectively privileged) institutional kinds, and that institutional kinds are functional kinds. This metaphysical thesis doesn't entail much about how institutions should be grouped and labelled. At best, this metaphysical thesis entails that insofar as one wants to take a scientific approach to institutions, one should label members of the same (objectively privileged) institutional kind with the same label. But this metaphysical thesis does not entail anything about what one ought to do if one also wants to promote social justice. So, on this interpretation, Haslanger's 'ameliorative' approach to social ontology is compatible with Guala's metaphysics of institutions, contrary to what Guala himself says.

This book is more than the sum of its parts. Instead of being a collection of stand-alone papers, it is an exploration of the central concepts that philosophers and social scientists have used to understand institutions, and it is a sustained exploration of how all these concepts hang together. Many of the chapters are so clearly and simply written that one could happily set them as readings for an introductory course. But, at the same time, the expert in social ontology will also find much that is instructive and provocative.

Christopher Clarke
Department of Philosophy
Erasmus University Rotterdam
clarke@fwb.eur.nl

References

Bacharach, M. [2006]: *Beyond Individual Choice,* Princeton: Princeton University Press.

Gilbert, M. [1989]: *On Social Facts,* Princeton: Princeton University Press.

Gold, N. and Sugden, R. [2007]: 'Collective Intentions and Team Agency', *Journal of Philosophy*, **104**, pp. 109-37.

Haslanger, S. [2012]: *Resisting Reality: Social Construction and Social Critique,* Oxford: Oxford University Press.

Hindriks, F. [unpublished]: 'Rules and Institutions: Essays on Meaning, Speech Acts, and Social Ontology', PhD Dissertation, Erasmus University Rotterdam

Lewis, D. [1969]: *Convention: A Philosophical Study,* Cambridge: Harvard University Press.

Morton, A. [1994]: 'Game Theory and Knowledge by Simulation', *Ratio,* **7**, pp. 14-25.

Searle, J. [1995]: *The Construction of Social Reality,* London: Penguin.