Précis of *Understanding Institutions*[1]

Francesco Guala

Università degli Studi di Milano, Italy

There is general agreement among social scientists that institutions are crucial determinants of economic growth and human flourishing. The consensus is that they are more important than natural resources: a well-organized group of people can prosper in a harsh environment, while badly organized societies usually go astray even in rich and generous ones. And yet, in spite of their importance, institutions are somewhat mysterious entities. What is an institution? If we do not know what institutions are, how can we possibly hope to improve their performance?

These are both philosophical and scientific questions. Philosophers have been asking “What-is-X” questions since the time of Socrates at least. Over the centuries however many of these questions have been taken over by science. If we want to know what is matter, or light, or life, we now ask physicists and biologists as well as philosophers. Similarly, questions about the nature of institutions cannot be answered satisfactorily without the help of science. So an important goal of *Understanding Institutions* is to offer a coherent picture of the fundamental architecture of modern societies, combining the insights of social scientists and philosophers who work on this topic. Another goal is to show that an adequate understanding of the nature of institutions helps resolve old conceptual and methodological problems in the philosophy of social science. While some of these problems simply disappear, others become more tractable once they are seen from the perspective of the unified theory.

In the course of the book I survey different views of institutions, analyze them critically, and explain how they relate to each other. I begin (in Chapter 1) by drawing a distinction between those theories that view institutions as rules, and those that view institutions as equilibria of strategic games. Then, I argue that these two approaches are complementary, and that they can be unified within a single framework.

The equilibria approach spans across the divide between philosophy and social science. The seminal theory in this tradition was proposed by David Lewis in a justly celebrated book on *Convention* (1969), but over the past four decades several other philosophers and social scientists have proposed equilibrium-based accounts of social institutions.[2] Theories within the equilibria approach view institutions as behavioral patterns that tend to persist because individuals have no incentive to deviate from the pattern unilaterally (unless everyone else does the same).

In spite of its explanatory achievements and its mathematical elegance, the equilibria approach has not been universally endorsed however. According to an equally popular alternative, institutions should rather be conceived as rules or constraints that guide the actions of individuals engaged in social interactions.[3]

[1] The papers that constitute this review symposium were originally presented at two conferences that took place in September 2017 – the European Network of Social Ontology in Lund (Sweden), and the European Philosophy of Science Association in Exeter (UK). The commentaries were later written up and submitted jointly to *Philosophy of the Social Sciences* by S.M. Amadae, who is the editor of the Symposium. I am extremely grateful to all the participants, and especially to Sonja, for the attention they have devoted to my book. Various paragraphs from the Preface and the Introduction of *Understanding Institutions* have been reproduced in this précis.

[2] See e.g. Lewis (1969), Ullmann-Margalit (1977), Sugden (1986), Skyrms (1996, 2004), Calvert (1998), Young (1998), Vanderschraaf (2001), Binmore (2005), Bicchieri (2006).

[3] Cf. e.g. Parsons (1935), Knight (1992), Mantzavinos (2001), Hodgson (2006), Miller (2010).

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Commento [MOU1]: Please designate in some fastion, “my,” “UI’s”, etc.

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Commento [MOU2]: Citations should be added here. Good ones, in addition to Margaret Gilbert, would be Gerald Gaus, Order of Public Reason, Joseph Heath, On Following the Rules

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Formattato: Tipo di carattere: +Titoli (Calibri)

The rules account is close to our vernacular, pre-prescientific understanding of institutions: intuitively, institutions regulate behavior, making certain actions appropriate or even mandatory in specified circumstances. The institution of private property, for example, regulates the use of resources by indicating who has access to them. The institution of money regulates the use of paper certificates in economic transactions. And the institution of marriage regulates the behavior of two or more individuals who pool their resources to raise kids, manage property, and help each other in many different ways.

But if institutions are rules, how do they influence behavior? Stating a rule is clearly insufficient to bring about an institution. To realize why, consider that there are plenty of ineffective rules: rules that are officially or formally in existence but that are nevertheless ignored by the majority of people. Traffic lights in Milan are regulation, in Rome they are a suggestion, and in Naples they are just decoration, as the saying goes. But since the rules are formally the same in Milan, Naples, and Rome, there must be something else going on. There must be some special ingredient that makes people follow the rules in some circumstances and ignore them in others.

The equilibria account of institutions tells us what the special ingredient is: effective institutions are backed up by a system of incentives and expectations that motivate people to follow the rules. An equilibrium in game theory is a profile of actions or strategies, one for each individual participating in a strategic interaction (Chapter 2). Each action may be described by a simple sentence of the form “do X” or “do Y.” The defining characteristic of an equilibrium—which distinguishes it from other profiles—is that each strategy must be a best response to the actions of the other players or, in other words, that no player can do better by changing her strategy unilaterally. If the others do their part in the equilibrium, no player has an incentive to deviate.

Since the actions of a strategic game can be formulated as rules, equilibrium-based and rules-based accounts of institutions are compatible. From the point of view of an external observer, an institution takes the form of a regularity that corresponds to the equilibrium of a coordination game. But each equilibrium strategy also takes the form of a rule that dictates each player what to do in the given circumstances. By combining the rules account with the equilibria account we obtain a unified theory that I will call the *rules-in-equilibrium* approach to the study of institutions.[4] Rules by themselves lack the power to influence behavior, but together with the right system of incentives and beliefs, they can influence the behavior of large groups of individuals. Institutions, in a nutshell, are rules that people are motivated to follow. Chapter 3 shows how this approach can be applied to the paradigmatic institution of money.

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Institutional rules sometimes simply state that we must “do X” or “do Y.” In many cases, however, they are conditional statements that prescribe different actions depending on the occurrence of certain events (“if X then do Y”). For example, the rules of traffic state that you must stop at the crossroads if the traffic light is red, proceed if it is green. Similarly, in many societies the actions of individuals are regulated according to their identities—there are rules of courtesy like “ladies first,” as well as hierarchical rules like “give orders if you are the husband, follow them if you are the wife.” Biological traits in such cases are used as signals that facilitate coordination, pretty much as traffic lights help us drive around smoothly. (If this statement sounds perplexing, let me clarify that these arrangements are not necessarily good equilibria: perhaps we would be better off if women gave orders and men obeyed; similarly, we could stop when the light is green and proceed when it is red.)

[4] Antecedents of this ‘hybrid’ approach can be found in Aoki (2007, 2011), Greif and Kingston (2011), as well as Guala and Hindriks (2015), Hindriks and Guala (2015).

Traffic lights and biological traits are correlation devices, and the actions of people who use these signals constitute correlated equilibria (Chapter 4).[5] Correlation devices multiply the number of ways in which we can try to coordinate. Suppose that you and I want to organize a dinner party. To simplify, let us suppose that we do not have strong preferences regarding the division of labor. To make sure that we coordinate, I text you a message that says: "I shop and you cook." The main purpose of this signal is to create the expectation that I will go shopping. Because if you believe that I will go shopping, then you will do the cooking, and the party will be a success. But of course this is just one of many possible signals that we could have used to coordinate. Had I told you "I cook and you shop," the opposite equilibrium would have been implemented. So language is a tremendously versatile device to create institutions, by sending signals that people use to converge on new equilibria. Humans are special in the animal kingdom in large part because they have language, and because they can use it to create a wide range of different social arrangements.

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

This point has not passed unnoticed of course. The most original and systematic attempt to place language at center stage in social ontology is the theory of constitutive rules proposed by John Searle (1995, 2010). Although this theory is a variant of the rule-based account of institutions, it attempts to explicate institutions using a very different kind of rule that, instead of merely regulating behavior, creates the possibility of new types of behavior. Constitutive rules according to Searle are statements of the form "X counts as Y in C," where Y denotes an institutional entity or fact or property, X is a preinstitutional entity, and C is a set of circumstances or conditions of instantiation. In the case of money for example a constitutive rule is: "Bills issued by the Bureau of Engraving and Printing (X) count as money (Y) in the United States (C)" (Searle 1995: 28).

Searle contrasts constitutive rules with~~to~~ regulative rules that have as their syntax "do X," or "if X do Y." The actions or strategies that appear in game-theoretic accounts of institutions have precisely this form, so Searle's distinction suggests that there is a deep hiatus between his own approach and the accounts of institutions found in the social science literature. But if this were true, then the attempt to unify different approaches to social ontology would fail: not all institutions would be systems of (regulative) rules in equilibrium.

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

There are good reasons, however, to believe that Searle's distinction between regulative and constitutive rules does not hold (Chapter 5). Using an argument originally devised by Frank Hindriks (2005, 2009), I show that constitutive rules have a much more limited role than the one envisaged by Searle: they are term-introducing principles that state the conditions of application of the theoretical terms that we use to label institutions. They are, first and foremost, naming devices for regulative rules.

The constitutive rule of money, for example, specifies the conditions that have to be satisfied for something to be money (it must be a paper bill issued by the Bureau of Engraving and Printing), and implicitly specifies what to do with paper certificates of that kind (use them to trade other commodities, save them for future purchases, etc.). Hindriks's view that regulative rules can be transformed into constitutive rules via the introduction of theoretical terms highlights the fact that constitutive rules do not add anything that cannot be expressed by means of simple regulative rules. In principle they could even be eliminated from our theoretical vocabulary, without causing any substantial ontological loss. The constitutive rule of money for example can be translated in a regulative rule such as: "if a bill has been issued by the Bureau of Engraving and Printing, then use it to purchase commodities or save it for the future," and so forth.

UI's~~The~~ unified theory thus helps attain ontological parsimony and at the same time offers an explanation of the pragmatic function of institutional terms (why they are useful and how they help us coordinate).

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

[5] The concept of correlated equilibrium is due to Aumann (1974, 1987). On its relevance for understanding conventions and institutions, see Vanderschraaf (1995) and Gintis (2009).

Chapter 6 illustrates what is the relationship between equilibria and functions, and explains how the normative (or 'deontic') power of institutions can be accounted for within the rules-in-equilibrium approach under the assumption that only incentives motivate action, as opposed to obligations, commitments, or other non-instrumental sources of normativity. The basic idea is that institutional rules create rights and obligations, specifying actions that can or must be performed in certain circumstances. These deontic powers may be represented as costs that transform individual incentives in strategic games. This modelling modeling strategy allows the extension of the unified theory to a wider class of games, including dilemmas of cooperation.[6]

Commento [MOU3]: This should be stated explicitly somewhere. You could consider citing:
Normativity and Instrumentalism in David Lewis' Convention
S. M. Amadae
History of European Ideas 37 (3):325-335 (2011)

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Chapters 7 and 8 (the 'Interlude') offer a speculative detour into the cognitive basis of institutions. The key challenge posed by the equilibrium approach is to explain how people achieve coordination in games with multiple equilibria (or multiple potential rules). To achieve coordination, people must build concordant expectations about each other's behavior. This process cannot be explained satisfactorily if we conceive of mindreading as a theoretical exercise. So, following Morton (2004) I suggest that we often coordinate *simulating* the mental processes of the other individuals with whom we interact.

Chapter 8 argues that the particular simulation process described by Morton offers a general template to understand other forms of coordinated behavior. The main skill for the creation of institutions is the capacity to identify a solution and to derive from it the actions to be performed by each individual from it. This kind of "solution thinking" can be carried out both in individualistic and in collectivistic modes, for example when individuals reason and act as members of a team.[7] However, contrary to what some philosophers have argued, many institutions do not require a joint intention or commitment to follow the rules.

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

The second part of the book is devoted to articulate the unified theory of institutions in more detail, and to explore its philosophical implications. In particular, I focus on the implications of the unified theory for the explanatory and predictive ambitions of social science.

For well over a century social scientists have been discussing the methodological foundations of their discipline. On the one hand, methodological "monists" have been arguing that the social sciences must follow the same approach as the natural sciences. On the other hand, methodological "pluralists" have argued that the very nature of social reality makes it impossible for social scientists to attain the same explanatory and predictive success of the natural sciences. Social scientists should adopt a different approach and give up the traditional goals of naturalistic scientific inquiry.

What ontological differences may license this kind of skepticism? A classic cause of concern has been the mind-dependence of social reality. The idea is that social entities differ from natural entities in that the former, but not the latter, depend essentially on our representations. The nature of a dollar bill, the fact that it is money, for example, depends on a collective belief or recognition that it is money—that it can be used to buy certain commodities and services. (Otherwise, it would be just a piece of paper with a picture of George Washington printed on it.) In contrast, a molecule of water is water regardless of what anybody believes about it. It does not have to be represented as water, in order to be what it is.

The thesis of mind-dependence has been used by many theorists to challenge the scientific ambitions of social science. The challenge can take different forms, however, depending on how the concept of dependence is interpreted. So part of the book is devoted to distinguish between different versions of the

[6] Cf. e.g. Crawford and Ostrom (1995), Bicchieri (2006).

[7] On 'team reasoning' see e.g. Sugden (2000), Bacharach (2006), Gold and Sugden (2007).

dependence thesis. In particular, it is useful to distinguish between causal and noncausal dependence on representations.

Commento [MOU4]: Please add one sentence explaining the essence of causal and noncausal dependence for readers less familiar with these issues.

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

I argue that the thesis of causal dependence is true, but that its philosophical consequences have been exaggerated (Chapters 9 and 10). Many social entities are involved in 'reflexive loops' with the categories that we use to classify them. This peculiar phenomenon can be captured by game-theoretic models where actions and beliefs sustain each other in equilibrium. Such models explain how a category that describes a behavioral regularity may contribute causally to the stabilization of that behavior.

An influential version of the causal dependence thesis has been proposed by Ian Hacking (1995, 1999), who has argued that social kinds differ from natural kinds because they are "interactive." Contrary to what he has claimed, however, Hacking's interactivity does not demarcate sharply between natural and social entities. Interactive kinds are as real as natural kinds, often support inductive inferences, and can be studied scientifically.[8]

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Thus mind-dependence, when it is interpreted causally, does not constitute a threat to the scientific ambitions of social science. The thesis of noncausal or ('ontological') dependence, in contrast, is a real threat for realism about social scientific kinds (Chapter 11). Such a thesis, if true, would demarcate between social and natural science. Under some readings, it would also imply anti-antirealism and infallibilism about social kinds: the properties of these kinds would not support inductive inference, but they could be known directly and without error by the members of the relevant community (Ruben 1989, Thomasson 2003).

The thesis of ontological dependence however is false (Chapter 12): any social kind may exist independently of anyone holding a correct theory of that kind. There is no guarantee, for example, that people understand what money is, or that the things that people classify as money actually are money. The nature of an institution is determined by its function, not by what people think about it. As a consequence, we ought to be realists and fallibilists about social kinds.

Commento [MOU5]: Isn't your argument stronger: "any social kind may exist independently of everyone holding a correct theory of that kind"?

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

The final two chapters are devoted to an issue that is currently hotly debated in many countries, concerning the design and identity of one of our most important institutions. The issue is whether to reform the institution of marriage so as to make it possible for partners of the same sex to get married. As we shall see, traditionalists have claimed that the institution of marriage is intrinsically or necessarily limited to heterosexual couples, and that the inclusion of same-sex couples would turn it into a different institution. The claim has sometimes been backed up by sophisticated semantic arguments, and philosophers have been engaged in the battle on both sides of the field (Chapter 13).

My own view is that it is perfectly legitimate to use the term "marriage" to refer to the contracts that regulate the relationships between individuals of the same sex. However, the debate on marriage highlights an interesting problem: it suggests that it is difficult to be simultaneously a realist and a reformist about institutions. Some philosophers have argued that the identity of institutions does not depends not on the rules that people actually follow, but on those that they should follow—that is, on the normative targets that we set for ourselves as a community. This "ameliorative" approach (Haslanger 2012) however is incompatible with realism. So I will propose a different solution based on the unified theory, to save both the realist principle that institutions do not depend noncausally on our intentions, and the reformist intuition that the rules of the game can be re-redesigned without changing the identity of an institution (Chapter 14). I argue that we can save realism and reformism by drawing a distinction between institutional types and tokens. While institution tokens are particular solutions to coordination problems, institution

Commento [MOU6]: Fix sentence

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Commento [MOU7]: Is it possible to say "depend causally" to avoid the double negative?

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

Commento [MOU8]: Can you make explicit, e.g.: "between institutional types and tokens by recognizing that social kinds relevant to institutions have the same ontological status as natural kinds"—something to this affect?

Formattato: Tipo di carattere: +Titoli (Calibri), 11 pt

[8] See also Mallon (2003), Khalidi (2013).

types are identified by their function, or the kind of strategic problems that they solve. For example, same-sex unions are marriages because they fulfill some of the classic functions of marriage.

References

Aoki, M. (2007) "Endogenizing Institutions and Institutional Change", *Journal of Institutional Economics* 3: 1-31.

Aoki, M. (2011) "Institutions as Cognitive Media between Strategic Interactions and Individual Beliefs", *Journal of Economic Behavior and Organization* 79: 20-34.

Aumann, R. (1974) "Subjectivity and Correlation in Randomized Strategies", *Journal of Mathematical Economics* 1: 67-96.

Aumann, R. (1987) "Correlated Equilibrium as an Expression of Bayesian Rationality", *Econometrica* 55: 1-18.

Bicchieri, C. (2006) *The Grammar of Society*. Cambridge: Cambridge University Press.

Binmore, K. (2010) "Game Theory and Institutions", *Journal of Comparative Economics* 38: 245-252.

Calvert, R. L. (1998) "Rational Actors, Equilibrium, and Social Institutions", in *Explaining Social Institutions*, edited by J. Knight and I. Sened. Ann Arbor: University of Michigan Press.

Crawford, S. E. S. and Ostrom, E. (1995) "A Grammar of Institutions", *American Political Science Review* 89: 582-600.

Gintis, H. (2009) *The Bounds of Reason*. Princeton: Princeton University Press.

Gold, N. and Sugden, R. (2007a) "Collective Intentions and Team Agency", *Journal of Philosophy* 104: 109-137.

Greif, A. and Kingston, C. (2011) "Institutions: Rules or Equilibria?" in *Political Economy of Institutions, Democracy and Voting*, edited by N. Schofield and G. Caballero. Berlin: Springer, pp. 13-43.

Guala, F. and Hindriks, F. (2015) "A Unified Social Ontology", *Philosophical Quarterly* 65: 177-201.

Hacking, I. (1995) "The Looping Effect of Human Kinds", in *Causal Cognition: A Multidisciplinary Debate*, edited by A. Premack. Oxford: Clarendon Press, pp. 351-83.

Hacking, I. (1999) *The Social Construction of What?* Cambridge, Mass.: Harvard University Press.

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

Hindriks, F. (2005) *Rules and Institutions: Essays on Meaning, Speech Acts and Social Ontology*. PhD Dissertation, Erasmus University Rotterdam.

Hindriks, F. (2009) "Constitutive Rules, Language, and Ontology", *Erkenntnis* 71: 253-275.

Hindriks, F. and Guala, F. (2015a) "Institutions, Rules, and Equilibria: A Unified Theory", *Journal of Institutional Economics* 11: 459-480.

Hodgson, G. M. (2006) “What Are Institutions?”, Journal of Economic Issues 15: 1-23.

Khalidi, M. A. (2013) *Natural Categories and Human Kinds*. Cambridge: Cambridge University Press.

Lewis, D. K. (1969) *Convention: A Philosophical Study*. Cambridge, Mass.: Harvard University Press.

Mantzavinos, C. (2001) *Individuals, Institutions, and Markets*. Cambridge: Cambridge University Press.

Miller, S. (2010) *The Moral Foundations of Social Institutions*. Cambridge: Cambridge University Press.

Morton, A. (2003) *The Importance of Being Understood*. London: Routledge.

Parsons, T. (1935) “The Place of Ultimate Values in Sociological Theory.” *International Journal of Ethics* 45: 282–316.

Ruben, D. H. (1989) “Realism in the Social Sciences”, in *Dismantling Truth*, edited by H. Lawson and L. Appignanesi. London: Weidenfeld and Nicolson, pp. 58-75.

Searle, J. R. (1995) *The Construction of Social Reality*. London: Penguin.

Searle, J. R. (2010) *Making the Social World*. Oxford: Oxford University Press.

Skyrms, B. (1996) *Evolution of the Social Contract*. Cambridge: Cambridge University Press.

Skyrms, B. (2004) *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.

Sugden, R. (1986) *The Economics of Rights, Co-operation and Welfare*. Oxford: Blackwell, 2nd edition 2004.

Sugden, R. (2000) “Team Preferences”, *Economics and Philosophy* 16: 174-204.

Thomasson, A. (2003) “Realism and Human Kinds”, *Philosophy and Phenomenological Research* 68: 580-609.

Ullmann-Margalit, E. (1977) *The Emergence of Norms*. Oxford: Clarendon Press.

Vanderschraaf, P. (1995) “Convention as Correlated Equilibrium”, *Erkenntnis* 42: 65-87.

Vanderschraaf, P. (2001) *Learning and Coordination*. London: Routledge.

Young, P. H. (1998) *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions.* Princeton:
Princeton University Press.