Social norms, expectations and sanctions: a comment on Frank Hindriks' essay

Francesco Guala

Università degli Studi di Milano

Abstract: Hindriks' paper raises two issues: one is formal and concerns the notion of 'cost' in rational choice accounts of norms; the other is substantial and concerns the role of expectations in the modification of payoffs. In this commentary I express some doubts and worries especially about the latter: What's so special with shared expectations? Why do they induce compliance with norms, if transgression is not associated with sanctions?

The amount of work that Frank Hindriks and I have done together makes it difficult to disagree on major issues. In 'Norms that Make a Difference' Hindriks however ventures beyond our Rules in Equilibrium (RiE) theory,[1] to explore an issue that we have skirted in our joint work. I have offered some tentative remarks on normativity elsewhere (Guala 2015, 2016), albeit in a slightly evasive way. In *Understanding Institutions* (2016) I have claimed that the normative power of institutions is unlikely to be backed up by a single mechanism. My argument is broadly evolutionary: norms are hugely important for the flourishing of our species, and we know that nature likes to select for the redundancy of important traits. So it is plausible that social norms are supported by a variety of different mechanisms.

This argument encourages complacency and procrastination, but even if it is many things, surely we have a duty to explain what normativity is and how it works. The rational choice approach gives an answer only to the second question (how it works), and Hindriks is not satisfied with it:

> Theories that analyse institutions as equilibria equate norms with sanctions and model them as costs. The idea is that the sanctions change preferences and thereby behavior. This view fails to capture the fact that people are often motivated by social norms as such, when they regard them as legitimate. I argue that, in order for a social norm to be perceived as legitimate, agents have to acknowledge reasons for conforming to it other than the sanctions they might incur for violating it. (Hindriks 2019: 125)

[1] See Guala & Hindriks (2015, 2019), Hindriks & Guala (2015, 2019).

The complaint is that standard rational choice accounts offer a limited view of normativity.[2] Hindriks argues that norms incentivize behavior in another important way, namely by *legitimizing* certain actions. A norm, in other words, does not simply tell you that certain bad consequences will follow if you don't comply; it also makes the complying attractive. It gives *positive* reasons to do something, rather than merely negative ones for not refraining from it.

I have two comments to make, one formal and one substantial. The two comments push in a certain direction – which is more friendly to rational choice accounts of norms than Hindriks would probably like. I merely offer them as stimuli to develop his argument further, however, for I suspect that Hindriks has more cards up his sleeve than he has shown so far.

I.

There is a frequent misunderstanding about the notion of 'cost', as it is used in rational choice models and in cost-benefit analysis more generally. It is standard practice in economics to focus on so-called *opportunity costs*, rather than on narrow monetary or material costs. The opportunity cost of an option is the value of what one misses by taking that option. For example: the opportunity cost of writing this commentary corresponds to the most valuable alternative way in which I could use my time right now. Suppose that the best alternative use of my time on this glorious day would be biking on the hills of Southern Piedmont. Obviously the cost of writing this commentary is greater, the more I like riding my bike.

The fundamental principle of rational choice theory is that people choose X over Y when the opportunity cost of doing Y exceeds that of doing X. So, as the opportunity cost of X increases, more and more people should switch from X to Y. If I found Hindriks' paper trivial and not worthy of critical attention, then today I would probably choose biking over writing.

This idea can be reformulated in the language of preferences: I choose X rather than Y if I prefer X to Y. So saying that a social norm decreases the cost of an option (e.g. X) is equivalent to saying that it makes Y more attractive compared to X. For example: I have promised the editor of *Analyse & Kritik* that I would write this commentary, and it is a social norm that promises must be kept. According to standard rational choice accounts, the norm about promising makes writing more attractive, by reducing the cost of writing. Without that norm, the opportunity cost of writing the commentary may be dangerously close to cost of the bicycle ride; but with the norm, it is reduced by the sanction that I would incur if I violated the norm.

[2] Crawford & Ostrom (1995) is a paradigmatic example, but see the other references in Hindriks' paper.

(The editor would spread the word that I am unreliable; my reputation would suffer; invitations to seminars and conferences would be withdrawn, etc.) This makes writing more attractive for me – its subjective value is more likely to exceed the value of riding my bike, given the promising norm; I am more likely to prefer writing, and eventually choose it.[3]

These preliminary remarks are meant to avoid unnecessary confusion. The confusion may arise because the technical notion of opportunity cost is not identical with the intuitive, everyday notion of cost. The latter, unlike the former, is tied to the idea of a *reduction* in the value of an option. Costs, according to everyday language, only subtract value. But addition and subtraction are not distinguished in the notion of opportunity cost. To reduce the relative value of Y (biking) is equivalent to increase the relative value of X (writing). To increase the value of X is to decrease its opportunity cost.

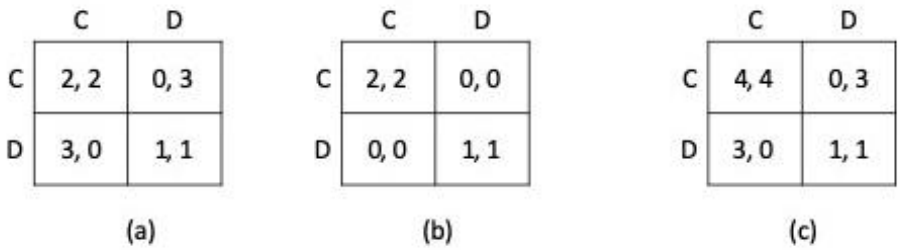

Figure 1: Transforming the Prisoner's Dilemma game

Since social norms are often meant to solve problems of cooperation, let us take a standard Prisoner's Dilemma (PD) game as an example, and let us examine two possible payoff transformations that may solve (or, better, 'dissolve') the problem. The PD game is represented in Figure 1(a), while (b) and (c) are possible solutions. The opportunity cost of choosing Defection is increased both in (b) and (c) by the payoff transformation.

Hindriks interprets the term 'cost' in the everyday sense. He takes equilibrium theorists to mean that norms always *decrease* the value of options by means of sanctions, as in solution (b). He then points out that this is not the only possible mechanism for the manipulation of payoff: it is also possible to *increase* the value of an option, as in solution (c).

[3] This presupposes an encompassing notion of preference. Preferences do not just reflect selfish motives, and do not just track our desires for material benefits. Not everyone agrees with this encompassing notion of preference, but I side with Dan Hausman (2011), who has provided the most comprehensive and persuasive analysis of this issue so far.

My first point thus is that the technical meaning of opportunity cost is compatible with Hindriks' proposal. Of course one may say: the technical meaning is less specific and hence less explanatory; the standard account overlooks an important distinction between causal mechanisms. I agree: I think that equilibrium models are great tools for representing problems of cooperation, but poor when it comes to identify the mechanics of decision-making. Hindriks' paper clarifies that there may be two ways to achieve the same goal (to create new equilibria, to sway people away from bad outcomes).

II.

But how would the new mechanism identified by Hindriks work? How can a norm enhance the value of an action or outcome? The key notion, according to Hindriks, is *legitimacy*. He says:

> A social norm exists when people believe that they are supposed to conform to its normative rule. Although sanctions can motivate people to conform, the norm as such can do so as well. In order for it to do so, people have to regard the norm as legitimate. This in turn means that they take the beliefs and expectations on which the norm depends to be justified. (Hindriks 2019: 144)

Notice that Hindriks's starting point is *perceived* legitimacy, i.e. people's belief that a norm is legitimate. But he goes on to suggest that legitimacy is provided by a set of expectations that are both 'justified and true':

> people can prefer to act on good reasons or, more precisely, to act on the basis of justified beliefs about what they ought to do. When they do, those beliefs increase the agent's payoffs for conforming to the norm. And they can do so to such an extent that they motivate the agent to conform. Given the proposed explication of the authority of social norms, this means that an agent follows such a norm exactly if: he prefers to conform to it because of a normative belief that is justified by empirical and normative expectations that are both justified and true. (Hindriks 2019: 141)

The challenge is to clarify what a 'justified normative belief' might be. Suppose that the norm governing behavior in a one-shot PD is: 'You ought to cooperate'. We know that there is no logical, knock-down argument that one ought to cooperate in a one-shot PD. If anything, there are strong arguments that one must *not* cooperate, based on dominance considerations. So a normative belief of this kind cannot be logically justified. To get around this, Hindriks argues that a rule can generate its own justification. How? The idea is that a normative belief is justified and true if and only if it is also widely shared in the relevant population.

> A natural suggestion is that a social norm exists exactly if it is generally known that people believe its normative rule applies. By this, I mean to say that they believe that the relevant activity is obligatory for those who participate in the social practice. In other words, the claim is that those who encounter a situation that constitutes a coordination game or a cooperation game are believed to be obligated to act in a particular manner and that this is generally known. (Hindriks 2019: 136)

Justification is 'bootstrapped', in other words, in a reflexive manner: legitimacy is provided by a shared belief that the norm is legitimate. This is, as far as I can tell, a coherent account of justification for normative beliefs. My worry however is: how does a system of shared expectations motivate people? What kind of extra reason could the fact that everyone believes that I ought to cooperate give me to engage in cooperation?

Notice that, according to Hindriks, a *personal* belief that I ought to cooperate has no legitimacy. A normative belief must be shared, in order to be 'justified and true'. He adds this requirement, I suspect, for two reasons: first, because he does not want every odd normative belief to count as legitimate; second, because personal whimsical beliefs are not genuine social norms. The sharing requirement excludes personal beliefs from the realm of social norms, and simultaneously provide legitimacy. For example: if Jill thinks that she ought to cooperate in a one-shot PD, but no one else does, then she is not under the influence of a social norm. Her normative belief, moreover, is arbitrary and unjustified — it lacks legitimacy in Hindriks' sense.

Still, this is independent of motivation. Jill's personal belief surely is motivating enough for her. So what sort of *extra reasons* do shared expectations provide? Doesn't a personal belief provide all the motivation that is needed to explain behavior? Perhaps Jill believes that cooperation is divinely mandatory, or that it is a Kantian categorical imperative. But the fact that she is mistaken should not make a difference to her payoffs, intuitively. Her *belief* that cooperation is legitimate should be enough to increase the value of cooperation. There doesn't seem to be any difference between a legitimate norm (in Hindriks' sense) and a norm that is believed to be legitimate by a single individual.

What does the fact that everyone else *also* believes that she ought to cooperate add to Jill's motivation, then? Accounts based on sanctions have a ready answer to this question: normativity is explained by our aversion to sanctions, or at least by an aversion to contradict others' expectations. This is exactly the 'cost' explanation that Hindriks would like to amend. It appeals to a familiar psychological and social mechanism

and does not require any notion of legitimacy. A sensitivity to others' beliefs and a capacity to anticipate the potential negative implications of norm-breaching is enough.[4]

Of course the sanctions account does not rule out that people also *endorse* social norms, or believe that they are legitimate. It is an empirical truth that social norms are usually approved by the members of the relevant community. But my belief that a norm is legitimate makes compliance more attractive regardless of the source of that belief, or whether anyone else thinks the same. It is the belief in legitimacy, rather than legitimacy itself, that does all the work.[5] If shared normative beliefs are important — and we all agree that they are — it is probably for some other reason. The standard account based on sanctions explains what that reason may be. I do not find a similarly persuasive story in Hindriks' essay, but I am confident that he has more to say and I am looking forward to hear about it.

References

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

Brennan, G./ P. Pettit (2004), *The Economy of Esteem*, Oxford

Crawford, S. E. S./E. Ostrom (1995), A Grammar of Institutions, in: *American Political Science Review* 89, 582–600

Guala, F. (2015), The Normativity of Institutions, in: *Phenomenology and Mind* 9, 116-127

— (2016), *Understanding Institutions*, Princeton

Guala, F./F. Hindriks (2015), A Unified Social Ontology, in: *Philosophical Quarterly* 65, 177–201

—/— (2019), Institutions and Functions, in: *Institutions in Action*, edited by Andina T./ P. Bojanic, Dordrecht

Hausman, D. M. (2011), *Preference, Value, Choice, and Welfare*, Cambridge

[4] Which explains its popularity: see e.g. Sugden (1998, 2000), Brennan and Pettit (2004), Bicchieri (2006).
[5] I'm rehearsing a familiar argument against 'true' normativity here, which has been developed by Turner (2010) in much more detail.

Hindriks, F. (2019), Norms that Make a Difference: Social Practices and Institutions, in: *Analyse & Kritik* 41, 125–145

—/F. Guala (2015), Institutions, Rules and Equilibria: A Unified Theory, in: *Journal of Institutional Economics* 11, 459–480

—/— (2019), The Functions of Institutions: Etiology and Teleology, in: *Synthese*, online first.

Sugden, R. (1998), Normative Expectations: the Simultaneous Evolution of Institutions and Norms, in: *Economics, Values, and Organization*, edited by Ben Ner, A./ L. Putterman, Cambridge

— (2000), The Motivating Power of Expectations, in: *Rationality, Rules, and Structure,* edited by Nida-Rümelin, J./ W. Spohn, Dordrecht

Turner, S. P. (2010), *Explaining the Normative*, Polity