Economics & Philosophy (2021), 1–18
doi:10.1017/S0266267121000195

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Institutions and their strength

Frank Hindriks

Department of Ethics, Social and Political Philosophy, University of Groningen, Oude Boteringestraat 52, 9712GL Groningen, the Netherlands
Email: f.a.hindriks@rug.nl

(Received 27 July 2020; revised 21 April 2021; accepted 27 April 2021)

Abstract

Institutions can be strong or weak. But what does this mean? Equilibrium theories equate institutions with behavioural regularities. In contrast, rule theories explicate them in terms of a standard that people are supposed to meet. I propose that, when an institution is weak, a discrepancy exists between the regularity and the standard or rule. To capture this discrepancy, I present a hybrid theory, the Rules-and-Equilibria Theory. According to this theory, institutions are rule-governed behavioural regularities. The Rules-and-Equilibria Theory provides the basis for two measures of institutional strength. First, institutions that pertain to coordination games solve problems of information. Their strength is primarily a matter of the expected degree of compliance. Second, institutions that concern mixed-motive games solve problems of motivation. Their strength can be measured in terms of the weight people attribute to its rule.

Keywords: Institution; normative belief; normative expectation; social norm; social practice

Introduction[1]

Strong institutions are important for a safe, prosperous and just society. In such a society corruption is low, investment is high, and people enjoy equal opportunities. Because of this, it is important to explain what makes an institution strong rather than weak. An account of institutional strength requires a theory of institutions. The most influential ones are equilibrium theories and rule theories (Rutherford 1994; Greif and Kingston 2011). According to equilibrium theories, institutions are stable behavioural regularities that solve coordination or cooperation problems.[2] A virtue

[1]This is the second part of a trilogy about norms and institutions. In the first, I introduce the Rules-and-Equilibria Theory, which explains how social norms can motivate in the absence of sanctions (Hindriks 2019). Here I use it to explicate the difference between strong and weak institutions. And in the third part, I discuss why and how equilibrium theories and rule theories should be combined or unified (Hindriks 2021).

[2]See instance Lewis (1969), Ullmann-Margalit (1977), Schotter (1981), Sugden (1986), Binmore (1994), Young (1998) and Binmore (2010).

of these formal theories is that they model the motivation of the participants of institutions in a *precise* manner. In contrast, rule theories regard institutions as rules that structure human interaction.[3] They lack the precision of equilibrium theories (Greif and Kingston 2011: 14). However, they provide a rich account of the *normative* dimension of institutions. Rules can justify behaviour and provide participants with reasons for action. And a rule as such can motivate compliance when people regard it as legitimate (Bicchieri 2006).

Ideally, a theory of institutions combines the best features of both kinds of theories. In section 1, I present a hybrid theory that invokes both equilibria and rules: The Rules-and-Equilibria Theory.[4] I argue that it is *precise* and does justice to the *normativity* of institutions. Because it is hybrid, it can capture discrepancies between rules and behaviours, which arise when institutions are weak. For instance, littering might be rampant even though a rule is in force that prohibits it. An equilibrium theory equates this institution with observed behaviour. In contrast, a rule theory identifies it with the rule that is violated. The Rules-and-Equilibrium Theory accommodates both dimensions and takes the institution to consist of a rule-governed behavioural regularity. Crucially, a rule can govern behaviour even when it is not complied with. This means that there might be norm-based pressures to behave in a certain way that fail to produce compliance.

In section 2, I investigate how institutional strength can be measured. Whereas most theories of institutions focus on one particular kind, I discuss both institutions that serve to coordinate behaviours and those that require cooperation. This is important in this context, because they concern different problems: the former present problems of information, the latter problems of motivation. And this has consequences for how their strength is to be measured. I propose that the strength of a coordinative institution is primarily a matter of the expected degree of compliance. In contrast, the strength of a cooperative institution can be measured in terms of the weight people attribute to its rule. Together with the Rules-and-Equilibria Theory, these measures constitute an important step towards understanding institutional strength and explaining what makes an institution strong rather than weak.

1. Institutions

1.1. Equilibrium theories

Theories of institutions can, roughly speaking, be divided into two kinds: equilibrium theories and rule theories (Greif and Kingston 2011). Insofar as they are economic theories, the latter are part of Old Institutionalism, whereas the former represent New Institutionalism (Rutherford 1994). Equilibrium theories originate with David Lewis' (1969) theory of conventions. Such theories identify institutions with stable behavioural regularities. Using the tools of game theory,

[3]Rule theories include Hart (1961), North (1990), Bloor (1997), Hodgson (2006) and Ostrom (2015).

[4]See Hindriks (2019) for an earlier version of this Rules-*and*-Equilibria Theory. It is closely related to the Rules-*in*-Equilibrium Theory, which I discuss in section 1.4 (Hindriks and Guala 2015; Guala and Hindriks 2015; Guala 2016).

they explain such regularities in terms of expectations about behaviours and preferences that are conditional on them. For instance, people typically want to match the way others greet them. Once a particular way of doing so is common, no one has an incentive to unilaterally deviate from the existing regularity. This means that the regularity is stable and that the behaviour, or the strategy players rely on, constitutes an equilibrium. Lewis was concerned with coordination problems, or situations in which people's preferences align. In addition to pure coordination games, they include games such as hi-lo, the battle of the sexes, and stag-hunt, in which payoffs differ between players or between equilibria. In such situations, people benefit from coordinating their behaviour with that of others. Solutions to coordination problems are conventions.

The idea that institutions are stable behavioural regularities has been extended to cooperation problems (Ullmann-Margalit 1977; Schotter 1981; Sugden 1986; Young 1998; Binmore 2010). Cooperation problems have a different structure: it is beneficial for all if everybody cooperates; however, each individual has an incentive to defect. Thus, they harbour a conflict of interest. Think, for instance, of a situation in which you can dodge the bullet when it is your turn to pay for a round of drinks, because people have not really been paying attention. Cooperation problems concern mixed-motive games such as the prisoner's dilemma, the ultimatum game and the trust game.

Equilibrium theories invoke norms in particular to explain how conflicts of interests can be resolved. They explicate norms in terms of sanctions, which are negative responses to rule violations such as fines or frowns. Players expect to incur sanctions with a certain probability. The expected costs of a sanction can change their behaviour, because they lower the payoff of defecting. Cooperation is secured if those costs are so large that the cooperation problem is thereby transformed into a coordination problem (Ullmann-Margalit 1977; Bicchieri 2006). Thus, equilibrium theories model norms as sanction-induced regularities in behaviour. I refer to this way of modelling norms as 'the norms-as-sanctions view.'[5]

Proponents of equilibrium theories commonly characterize institutions as solutions to cooperation and coordination problems, in which case they require compliance (Lewis 1969; Schotter 1981). In this vein, Guala maintains that institutions are effective rules, which are 'rules that people are motivated to follow' (Guala 2016: xxv). The problem with this is that, by requiring compliance, such theories exclude weak institutions. As I discuss in more detail in section 2, coordination institutions will be weak, for instance, when the (expected) degree of conformity is low. Cooperation institutions are particularly sensitive to the expected cost of sanctions, which depends on their probability and severity. If it is too low, they are likely to be weak. Furthermore, the probability with which norm violations are detected is influenced by the costs of observing them. These can be high, especially if perpetrators try to conceal

[5]Lewis (1969: 97–100) argued that conventions are norms. More recently, people have argued that this need not be the case (Bicchieri 2006; Brennan *et al.* 2013; Hindriks 2019). In contrast to a cooperation norm, a coordination norm does not induce a behavioural regularity but reinforces it.

them, as is typical for people who lie or commit adultery. Factors such as these can account for weak institutions within an equilibrium framework.[6]

Theories that do not feature rules can be used to model social practices, or so I propose. In contrast to institutions, social practices do not involve norms (Tuomela 2002). I propose that a social practice is a regularity *R* in some population *P* and a situation *S*, which is either (a) a coordination game or (b) a mixed-motive game [PRACTICE]:[7]

> [PRACTICE] A social practice *R* in a recurring situation *S* within population *P* exists exactly if enough members of *P*:
>
> 1. expect everyone to conform to *R*,
> 2. prefer to conform to *R* (a) conditional on 1 or (b) unconditionally.

This definition entails that social practices are stable interdependent regularities. Social practices can be (a) conventions, such as driving on the right, greeting practices and dialects. But they can also be (b) competitive practices, such as overfishing, an arms race, or a rat race at the workplace. As these examples reveal, they can be local (dialects), specific to a society (in Japan, people greet by bowing), or even more widespread (driving on the right).[8]

Especially because it can easily be formalized, PRACTICE offers a good basis for modelling people's incentives in a precise manner. This will turn out to be particularly useful in section 2, where I develop two measures of institutional strength. However, as I go on to argue, rule theories provide for a better way of understanding the normative dimension of institutions norms-as-sanctions view.

1.2. Rule theories

Douglas North (1990: 3–5) characterizes institutions as the rules of the game that define the way it is played. By doing so, they provide a conception of what to do in a particular situation. North also claims that institutions are rules or constraints that structure human interaction. People meet such standards of behaviour because they

[6]I thank an anonymous referee for pressing me to discuss some of the tools that equilibrium theory has to capture weak institutions. Becker (1968) already argued that people will break laws when the expected utility of the action outweighs the expected cost of the punishment. Insofar as informal norms are concerned, Keuschnigg and Wolbring (2015) argue that people rely on casual observation concerning norm violations to estimate the expected costs of sanctions. When people perceive some form of social disorder, they infer that those costs are low. This in turn leads to an increase in norm violations for minor transgressions.

[7]Conditions 1 and 2 imply the existence of regularity *R*. Because of this, I do not include its existence as a separate condition, as Lewis (1969) does. Another difference is that I do not require common knowledge (Binmore 2008). The number of members of *P* that satisfy conditions 1 and 2 is high enough when conforming to *R* forms a Nash equilibrium.

[8]According to the Folk Theorem in game theory, cooperation can be an equilibrium in an infinitely repeated mixed-motive game. However, in practice cooperation rarely lasts long if there is no social norm to sustain it (Fehr and Gächter 2000*a*, 2000*b*; Bicchieri 2006). Because of this, the analysis abstracts from the possibility of cooperative practices.

might be punished if they violate them. Thus, rule theories also invoke sanctions to explain behaviour. And they are motivated to conform to a rule if and because it is enforced it (North 1990: 4 and passim; Greif and Kingston 2011: 40).

Rules constrain behaviour by limiting the options people have. However, they also enable certain actions or outcomes. Hodgson makes the point as follows:

> The existence of rules implies constraints. However, such a constraint can open up possibilities: it may enable choices and actions that otherwise would not exist. For example: the rules of language allow us to communicate; traffic rules help traffic to flow more easily and safely; the rule of law can increase personal safety. Regulation is not always the antithesis of freedom; it can be its ally. (Hodgson 2006: 2)

In light of this, Hodgson defines institutions as 'systems of established and embedded social rules that structure social interactions' (Hodgson 2006: 18).

According to some rule theories, normative rules can motivate people directly. A normative rule presents some kind of action as obligatory ($\mathfrak{R}$). Using '*C*' for a kind of action, its basic structure is: 'In *S*, it is obligatory to do *C*'.[9] An individual is motivated directly by a normative rule when she recognizes the force of the obligation and this favourably inclines her to conform. Someone who appreciates the force of an obligation will tend to feel guilt or shame when flouting it. The fact that rules as such can motivate serves to capture the Wittgensteinean notion of following a rule (Schatzki 1996; Bloor 1997). Someone follows a rule exactly if she conforms to it because of the rule (Brennan *et al.* 2013). Insofar as normative rules are concerned, this means that she conforms because she regards the rule as legitimate.

Hart (1961) criticizes the norms-as-sanctions view because it fails to capture this feature of social rules. Social norms cannot be reduced to the motivating effect of enforcement, because they can motivate directly. Hart argues that an agent who accepts a rule thereby regards it as legitimate. Furthermore, someone who accepts a rule treats it as a basis for justifying and criticizing conduct and regards it as a reason for action.[10] According to this Acceptance Theory, a social norm is a generally accepted normative rule. It entails that sanctions are not always required for people to comply with a rule. And it thereby reveals that the norms-as-sanctions view leaves something to be desired.

However, the Acceptance Theory suffers from a significant problem: its requirement of acceptance is too strong. Someone who accepts a rule believes in the rule: she believes that she ought to act accordingly. But this is not required for a rule to exist. A normative rule can in fact be in force without people subscribing to it. It suffices if people believe that they are supposed to comply. Consider someone who dislikes greeting acquaintances in public places such as streets and supermarkets and regards it as a silly practice. Even though he does

[9]Normative rules can also permit or prohibit certain actions, or feature any other first- or second-order right or obligation.

[10]By accepting a rule, an agent adopts what Hart (1961) calls 'the internal point of view'. See Shapiro (2006) for more on this.

not subscribe to the local greeting norm, he is distinctly aware of the fact that he is supposed to greet others, given that others accept the norm. And this could influence his behaviour. Now, it might be that everybody in the population merely expects others to accept the rule without doing so themselves. This would suffice, I propose, for it to be in force. In fact, social norm can even be practiced when everybody mistakenly expects the others to accept its rule.[11]

The point can be made more precisely by distinguishing between normative beliefs and normative expectations. The content of a normative belief is a normative rule: it is the belief that $\mathfrak{R}$. Someone who has a normative belief thereby accepts the relevant rule.[12] In contrast, a normative expectation. This is a higher-order expectation, the expectation that others believe that $\mathfrak{R}$ (cf. Bicchieri 2006, 2016). Someone who merely has a normative expectation, I will say, acknowledges the relevant rule. Now, for a rule to be in force, people must have a normative expectation, but not a normative belief. Hence, *pace* Hart (1961) and Brennan *et al.* (2013), acceptance is not required, whereas acknowledgement is. It follows that, according to what I call 'the Acknowledgement Theory', a social norm is a generally acknowledged rule.

In light of this, I define the notion of a social norm as follows [NORM]:[13]

> [NORM] A social norm $\mathfrak{R}$ exists in a population exactly if a substantial number of members of *P* have the normative expectation that $\mathfrak{R}$.

People who have normative expectations expect others to believe that they ought to act in a certain way. In more colloquial terms, they believe that they are supposed to perform that action.

1.3. The Rules-and-Equilibria theory

Institutions involve social practices as well as social norms. But how are they connected? It will not do to require norm-compliance, because weak institutions are institutions too. I propose instead that an institution is a social norm that governs a social practice. This means that the social norm makes the participants in the social practice more motivated to comply with it. In other words, the social norm increases the agents' payoffs for conformity. But it does not necessarily change how they rank the options. Hence, a social norm can govern a social practice even though the latter does not correspond to the former. This implies that a rule $\mathfrak{R}$ can govern the corresponding regularity *R* as well as an alternative regularity *R*'. Think, for instance, of an area where monogamy is popular but cheating even more so.

As discussed, a social norm can influence someone's motivation to comply to a norm directly, by increasing the payoff for conforming to it. It can also do so indirectly through sanctions that decrease the payoff for violating it. I will say

[11]In such situations, they exhibit pluralistic ignorance (Prentice and Miller 1993; Kuran 1995).

[12]What I call 'a normative belief' can be interpreted as a genuine belief or in terms of some non-cognitive attitude (Brennan *et al.* 2013).

[13]'A substantial number' is vague, but in a way that suits the topic in general and the goal of investigating the strength of norms and institutions in particular.

that, in both cases, the individual attributes weight to the norm. This notion of the weight of a norm can be used to define that of norm-governance [GOVERN]:[14]

> [GOVERN] A social norm $\mathfrak{N}$ governs a social practice *R* or *R'* exactly if a substantial number of participants attribute non-trivial weight to $\mathfrak{N}$.

Together with PRACTICE and NORM, GOVERN defines the notion of an institution as a norm-governed social practice. GOVERN connects the equilibrium theory of social practices with the Acknowledgement Theory of social norms. Crucially, this proposal implies that a social norm can govern a social practice that does not correspond to it.

To see how all of the elements relate to one another, it is useful to present a more detailed analysis. As before, *S* is an interdependent situation that can be represented as a cooperation game or a mixed-motive game. *C* stands either for coordinate or for cooperate; *D* for deviate or defect. When people do *C* this results in regularity *R*; when they do *D*, this results in regularity *R'*. Given these stipulations, the notion of an institution can be analysed as follows [INSTITUTION]:

> [INSTITUTION] An institution $\mathfrak{N}$, according to which it is obligatory to do *C* in *S*, exists in a population *P* exactly if a substantial number of its members:
>
> (1) expect a number of others to believe that $\mathfrak{N}$,
> (2) expect some to do *C* and others to do *D*,
> (3) (a) believe that $\mathfrak{N}$ and/or (b) expect some to be disposed to sanction violations, at least in part because of 1,
> (4) are more favourably inclined to conform to $\mathfrak{N}$ because of 2, 3a and/or 3b, and
> (5) either doing *C* or doing *D* constitutes an equilibrium in *S*.

Condition 1 entails the existence of a social norm, condition 5 that of a social practice, and conditions 3 and 4 capture how the former governs the latter. Thus, the proposal explicates the conception of an institution as a norm-governed social practice.[15]

As discussed, a social norm governs a social practice when it increases people's motivation to conform to it. It does so directly when preferences are conditional on

[14]For simplicity, GOVERN mentions only two possible regularities. In coordination games, there can be more equilibria. Furthermore, regularities need not be uniform if there are differences between individuals such that some do not conform to them.

[15]A population can be a social group, a society or even all of humanity. Because of this, INSTITUTION applies at any level of generality. However, institutions at different levels can conflict. For instance, teenagers might encourage each other to damage property by means of graffiti even though there is a societal norm against doing so. Furthermore, institutions can require participants to differentiate between groups. There might, for instance, be a norm that Francophone retailers in the Canadian province of Quebec speak French to Canadians from other provinces but English to Americans. Finally, there might be an ingroup the members of which support a norm to discriminate against an outgroup. The proposed analysis can be developed further so as to take these complications into account. I thank an anonymous referee for pressing me to remark on these important issues.

normative beliefs (condition 3a).[16] And it does so indirectly when preferences are conditional on expectations about sanctions (condition 3b). A social norm can motivate individuals in both ways. Obviously, people are motivated by a norm or its sanctions only if they believe that the relevant norm exists. Because of this, conditions 3a and 3b are conditional on condition 1. Furthermore, a norm can motivate an agent without changing her behaviour. It then affects her payoffs without changing her ranking of the available options. To allow for this, condition 4 states that norms or sanctions more favourably incline people to conform. Finally, condition 2 specifies empirical expectations, which are mixed when people expect a certain proportion to do *C* and others *D*. Together, the expectations, beliefs and preferences that people have settle whether *R* or *R*' is the equilibrium or social practice (condition 5).

Because of the way in which it combines rule theories and equilibria theories, I call this 'the Rules-and-Equilibria Theory' (*RaE*) of institutions. In comparison to rule theories, *RaE* is more precise in particular in how it models motivation to comply with institutions. One of two advantages that this hybrid theory has over pure equilibrium theories concerns their normativity. The theory captures the motivating power of social norms along with the legitimacy they can be taken to have. Before discussing the second advantage, I need to say more about the relation between institutions and social norms. According to *RaE*, an institution is a social norm that governs a social practice. But a social norm need not govern a social practice in order for it to exist. In fact, it can exist without having any effect on the members of the population to which it applies. All that is required is that they have normative expectations (NORM). In terms of GOVERN, a social norm constitutes an institution exactly if it non-trivially increases people's motivation to comply. When it does not have such an effect or it is negligible, it is merely a social norm. In terms of INSTITUTIONS, a social norm can exist even though conditions 2–4 are not met. Thus, the notion of an institution is more demanding than that of a social norm.

The second advantage that *RaE* has over pure equilibrium theories: it provides for a richer account of weak institutions. Equilibrium theories equate institutions with behavioural regularities. The regularity that constitutes a weak institution differs from that of a strong institution. Hodgson (2006) argues that institutions are more than behaviours because they can continue to exist even when people stop acting as they should. Such discrepancies reveal that sometimes participants in an institution fail to meet a standard, an idea that is at best implicit in equilibrium theory. This can be accommodated by analysing institutions not only in terms of regularities but also in terms of rules. The conception of an institution as a social norm that governs a social practice does exactly that.

[16]When a norm motivates directly, the agent's normative belief increases the payoff of compliance. Here are four hypotheses about how it can do so. First, beliefs might motivate. Second, what I call 'a belief' is a non-cognitive attitude that motivates as such (see note 12). Third, individuals have a desire for acting according to their normative beliefs. Finally, they might have a 'desire to fulfill others' legitimate expectations' (Bicchieri 2006: 42).

1.4. A comparison

The Rules-and-Equilibria theory (*RaE*) has close affinity to two other hybrid theories of institutions.[17] The first is the Rules-*in*-Equilibrium theory (*RiE*) of Francesco Guala and Frank Hindriks (Hindriks and Guala 2015; Guala and Hindriks 2015; Guala 2016). Its main innovation is that it explains coordination in terms of signalling rules. When people face coordination problems, they rely on correlation or signalling devices, which are public signals that correlate their expectations (Aumann 1974; Gintis 2007). Such devices feature in signalling rules. Using 'Σ' for signalling devices and 'C' for a coordinative action, the basic structure of a normative signalling rule is: 'In S, it is obligatory to do C if Σ.' They enable people to form reliable expectations about each other's behaviour. As their preferences are conditional on such expectations, they will then adjust their behaviour accordingly.

The Rules-and-Equilibria theory accommodates the notion of a signalling rule. It can in fact be seen as an extension of *RiE*. First, whereas *RiE* models norms as costs, *RaE* offers a richer account of norms that captures the idea that rules can motivate as such. Second, *RiE* is restricted to coordination problems This is why, as its name reveals, rules are in equilibrium. In contrast, *RaE* extends to cooperation problems. And it allows for equilibria to come apart from the rules that govern them. For these reasons, it is not only richer, but also has a larger scope.

RaE also has affinity with Bicchieri's (2006, 2016) Conditional Preference theory (*CPT*) of social norms. Both theories feature social norms in terms of empirical expectations, normative expectations, sanctions and preferences that are conditional on some or all of these factors. However, *CPT* applies first and foremost to cooperation problems. Furthermore, *RaE* is more demanding. According to *CPT*, a social norm exists exactly if people have preferences that are conditional on empirical and normative expectations. This means that their preferences are such that they would conform if they were to have the relevant expectations. They do not need to have those expectations in order for the norm to exist. In fact, people who actually have them will comply with the norm (Bicchieri 2006: 11). In contrast, *RaE* requires normative expectations for a social norm to exist. As discussed, this entails that they believe that they are supposed to conform to the norm. Empirical expectations are needed for people to comply.

A second respect in which *RaE* is more demanding than *CPT* concerns the motivating power of social norms. *CPT* explicates it in terms of preferences that are conditional on normative expectations. If this is what motivates them, they are concerned with acting according to the beliefs of others. Bicchieri seems to recognize this when she observes that people who have normative expectations might act accordingly because the 'feel great social pressure' (2006: 14). *RaE* requires that preferences are conditional on normative beliefs as well. Because they must have those beliefs as well, they subscribe to the norm and take themselves to be bound by it. This means that he believes that he is obligated to perform the relevant action. In one respect, *RaE* is less demanding

[17]See Pettit (1995, 2007), Aoki (2001) and Greif and Kingston (2011) for other hybrid theories of institutions. In Hindriks (2021), I compare Pettit's (1995, 2007) Virtual Control Theory (*VCT*) to *RaE*. I argue that *VCT* accounts for strong institutions, but fails to accommodate weak ones.

than *CPT*. INSTITUTION requires that a social norm affects people's motivation, not necessarily their behaviour. As condition 5 has it, the norm favourably inclines them to conform. Because of this, the analysis explicitly accommodates the possibility that a practice diverges from the norm that governs it. Thus, only INSTITUTION captures the idea that a social norm can govern a social practice that does not correspond to it.[18]

The upshot is that *RaE* provides a systematic account of both cooperation and coordination institutions. It captures the *normativity* and *strength* of institutions in a *precise* manner. Because of this, INSTITUTION forms a suitable basis for the measures of institutional strength that I go on to develop.

2. Strength

Understanding strength is important because institutions can be valuable. There are two sides to this. When an institution is in fact valuable, strength is a virtue. Because of this, it is important to know which factors make it strong such that they can be promoted. However, institutions can also harm, discriminate, dominate, exploit and oppress people (Young 1990; Cudd 2006; Haslanger 2012). Such effects can occur within a group of participants or between different social groups. For instance, a racial group can develop a norm to disparage members of another racial group in order to increase its own status (Ellickson 1991; McAdams 1995). When institutions are objectionable in some such way, strength is a vice. A more precise appreciation of what makes institutions weak is vital for understanding social change.[19]

Thus far, I have emphasized the role of motivation in relation to institutional strength. But this is only part of the story. An institution is weak when compliance is low. However, an institution that is generally complied with need not be strong, as it could be on the verge of collapse. Inspired by Guala (2016: xxv), who characterizes institutions as 'effective rules', I will say that such an institution is 'effective.' A strong institution differs from an effective one in that compliance is robust. This means that people would still conform to it under more challenging conditions. Presumably, this is why Pettit claims that 'the investigation of resilience is vital for choosing policies and designing institutions' (Pettit 2007: 83). Robustness forms a safeguard against a breakdown, and often against a loss of value.[20]

The strength of an institution is determined by a wide range of factors. Think, for instance, of the leadership qualities of a president, the financial reserves held by banks, the racist or sexist stereotypes that prevail in a society, and the sermons people hear in church. What I am looking for, however, are generic factors that apply to all institutions. The two key factors, I propose, are motivation and

[18]For a more elaborate comparison between *RaE* and *CPT* see Hindriks (2021).

[19]For rational choice perspectives on social change, see Knight (1992), Bicchieri (2016) and Sunstein (2019). See Sankaran (2020) for an overview of theories of ideology critique as a means to social change.

[20]Pettit (2015) argues that robustness is valuable as such. The idea is that protection against a breakdown is valuable irrespective of a pending threat, just as a secure relationship is valuable even if it is never put to the test.

information. As just discussed, a cooperation problem is typically solved by means of a social norm that changes people's motivation. That norm modifies the situation such that interests no longer conflict. Information is key to solving coordination problems. They feature multiple equilibria. For instance, people greet each other in many different ways including giving a handshake or saying 'namaste'. Information about what others do is all that someone needs in order to successfully coordinate with others. Because of this difference, I develop two measures of institutional strength, one for coordination institutions and one for cooperation institutions.

2.1. Cooperation institutions

As an example of a cooperation institution, consider littering practices in Singapore.[21] People used to dispose of chewing gum left and right. They would leave it on pavements and on seats of public buses. And they would stick it in mail boxes, on elevator buttons and inside keyholes, which caused maintenance problems in high-rise public-housing apartments. The last straw was when vandals stuck gum on the door sensors of the Mass Rapid Transit trains that started operating in 1987, which disrupted services. This led to a ban on chewing gum in 1992, along with hefty fines. Given that it is no longer for sale in Singapore, it is primarily tourists who incur fines of $1,000. Singapore is now known as 'the Fine City', not only because of these fines but also because of its immaculate state. Littering has become rare more generally. Although this was initially because of the sanctions, Singaporeans have come to take pride in keeping the city clean. An increasing number of them will remind a family member not to litter when they notice them doing so. In a study released in 2019, 96.3% of respondents agreed that visitors admire the city's cleanliness.

Three things are particularly striking about this example. First, the equilibrium has changed from littering to non-littering. Second, the new equilibrium appears to be rather stable. And third, during the process, motivation has transformed from predominantly indirect to primarily direct. Bicchieri (2006) argues that this transformation is typical. Sanctions are the prime source of motivation for conformity in the beginning. Later people comply because they perceive the norm as legitimate. Presumably there was an informal norm back when littering was rampant. If so, the institution used to be weak. In fact, violations were so widespread that the equilibrium diverged from the norm. But now it is strong.

The strength of cooperation institutions, I propose, turns on the motivation of their participants, as explicated by condition 5 of INSTITUTION. Cooperative institutions are effective when cooperation is a minimally stable strategy, which means that they are motivated to cooperate but barely so. Strong and weak institutions can be characterized in terms of the notion of a motivation surplus and that of a motivation deficit. The participants in an institution have a motivation surplus (deficit) when they have more (less) motivation to cooperate

[21]See https://www.goabroad.com/articles/study-abroad/singapore-laws-to-know-before-you-go and https://www.businessinsider.sg/singaporeans-take-pride-in-keeping-the-city-clean-and-more-than-half-will-tell-those-around-them-not-to-litter-survey-finds/.

than needed for compliance. An institution is strong when there is a substantial motivation surplus. Many of its participants will continue to cooperate even if they face pretty strong pressures to violate the norm, which means that compliance is robust. A cooperation institution is weak when there is a significant motivation deficit. They then need strong additional incentives to start cooperating.

The notions of a motivation surplus and a motivation deficit can be used to develop a measure of the strength of a cooperative institution. Mixed-motive games harbour a conflict of interests. The payoff of cooperation, given that the others cooperate as well, is '*c*.' The payoff of unilateral defection is '*d*', with $d > c$. Because of this, players have an incentive to defect. But if they both do so, the resulting outcome is the worst possible for each. The prisoner's dilemma exhibits this structure. In a single-shot game, the participants will defect. In a repeated game, a social norm can enable cooperation. It does so when players attribute enough weight to it. The weight of a social norm can be defined in terms of so-called 'delta-parameters' (Crawford and Ostrom 1995). I use one for direct motivation, which is due to the norm itself (δ_n), and one for indirect motivation, which is due to sanctions (δ_s). The former represents the extent to which the norm increases the payoff of conforming ($c + \delta_n$); the latter the extent to which the sanction decreases the payoff of not conforming given the probability of being sanctioned ($d-\delta_s$).

As proposed above, full compliance is required for an institution to be effective. Now, in order for participants to be indifferent between cooperating and defecting, the following condition has to be met: $c + \delta_n = d-\delta_s$. The two delta parameters represent what I call 'the weight of the norm' (w): $w = \delta_n + \delta_s$. Participants are indifferent when this weight equals $d-c$. In order for cooperation to be a minimally stable strategy, it has to be slightly higher than this. What I call 'the compliance weight' of the norm is: $w_c = d-c + \epsilon$. An institution is effective precisely if the weight that a norm actually has (w_a) is equal to its compliance weight: $w_a = w_c$. When this is the case, the norm resolves the conflict of interests that is inherent to cooperation games, because defection ceases to be attractive.

The notion of compliance weight can be used to define those of a motivation surplus and a motivation deficit. There is a surplus when the actual weight is larger than the compliance weight: $w_a > w_b$. When it is lower, there is a motivation deficit: $w_a < w_b$. In light of this, I propose to measure the strength of a cooperative institution in terms of the following ratio (with 'SCOOP' for the strength of a cooperative institution):

$$\text{SCOOP} = w_a/w_c$$

This ratio equals 1 for effective institutions; it is (substantially) higher than 1 for strong institutions; and it is (substantially) lower than 1 for weak institutions. Given that $w_a = \delta_{na} + \delta_{sa}$ and $w_c = d-c + \epsilon$, this is equivalent to: $(\delta_{na} + \delta_{sa})/(d-c + \epsilon)$. Thus, the proposed measure is a function of the payoffs for cooperating and defecting and of direct and indirect motivations of the cooperative norm. SCOOP measures the relative distance between the weight of the norm and the point at which it tips the balance in favour of conformity.

By way of an example, consider cheating on an exam. Students prefer cheating to not cheating: $d = 20$ and $c = 10$. They are indifferent when the weight of this

norm is: $w_c = \delta_{nc} + \delta_{sc} = d{-}c = 10$. Suppose that an honour code is introduced that prohibits cheating along with appropriate sanctions. Students care quite a lot about sanctions and only a little about the code as such: $\delta_{as} = 4$, $\delta_{an} = 2$. In that case, they attribute the following weight to it: $w_a = 6$. The strength of the honor code at their university is $s = 0.6$. This means that the students suffer from a motivational deficit. Norm-violation will be widespread. Hence, the institution is rather weak. As it happens, students at another university are in the exact same situation except that they are substantially more moved by the thought that cheating is wrong, such that $\delta_{an} = 8$. The total weight of (this part of) the honour code at their university is $w_a = 12$. This means that the strength of their institution is $s = 1.2$. In this case, the institution is fairly strong. Students robustly comply with its norm. In this way, the proposed measure captures the strength of cooperation institutions.

2.2. Coordination institutions

According to a famous Italian proverb, traffic lights are instructions in Milan, suggestions in Rome and decorations in Naples (Guala 2016: xxiv). I take this to imply that that traffic practices approximate traffic rules in Milan, whereas they frequently deviate from them in Rome and even more often in Naples. This in turn suggests that traffic rules are stronger in the north as compared to the middle of Italy and they are virtually non-existent in the south. It is tempting to explain this in terms of the motivation of traffic participants. The different degrees of conformity would then reflect the fact that in Milan people are much more committed to following traffic rules than people in Rome, while in Naples they are not committed to them at all. However, it is not obvious that this is correct. Traffic lights resolve coordination problems, which turn on information rather than motivation.

In light of this, I propose to measure the strength of coordination institutions in terms of empirical expectations. Condition 2 of INSTITUTION allows for players to expect partial conformity. The idea is to measure the strength of a convention in terms of the expected degree of conformity (d_c). The baseline is the tipping point, the expected degree of conformity at which people switch from one equilibrium to another (d_{ct}). In these terms, the proposal for a measure of the strength of a coordinative institution is as follows (with 'SCOOR' for the strength of coordination institutions):

$$\text{SCOOR} = d_{ca}/d_{ct}$$

When this ratio equals 1, the coordinative institution is effective. And when it is substantially higher (lower) than 1, it is strong (weak).

An important feature of conventions is that they are self-enforcing. This supports a particular dynamic for empirical expectations. When SCOOR is larger than one, it will tend to increase. And when it is smaller than one, it will in all likelihood decrease such that the practice goes out of existence, although this can take some time. This suggests that there are few conventions that are moderately strong or weak. Consider Sweden's 1967 switch from driving on the left to driving on the right. Driving on the left caused a fair number of traffic accidents

for two reasons. First, because of the large amount of cross-border traffic from neighboring countries where they drove on the right. Second, because the vast majority of cars in Sweden were left-hand drive vehicles. The change was widely unpopular, which made the political process far from easy. Even so, the changeover went relatively smoothly. The number of accidents actually went down for a period of six weeks.

To effect the change, practical obstacles had to be overcome, including changing traffic signs and relocating bus stops. However, the crucial step was an information campaign backed by the authorities. By changing their empirical expectations, it enabled the Swedes to move away from the suboptimal equilibrium they were locked into. Within less than a day, the Swedes changed their behaviour from almost full compliance to one rule to almost full compliance to its alternative. In light of this, Bicchieri observes that 'a governmental diktat may easily work for conventions (such as traffic rules)' (2016: 144). She adds that, in such cases, 'the law coordinates behaviour via the creation of new expectations' (Bicchieri 2016: 144).

Even though coordination institutions are problems of information, they also have a normative dimension. Because coordinative regularities are self-enforcing, social norms merely add to the strength of a convention. They reinforce or stabilize them. Consider a pure coordination game in which the payoff for coordinating is 'c' and that for deviating from the existing regularity is 'd' with $d = 0$. A coordination norm makes coordinating more attractive. The extent to which it increases motivation is given by the weight people attribute to its norm, $w = \delta_n + \delta_s$. The extent to which a social norm contributes to people's motivation to conform is w/c. Such motivation might make people more careful, enhance their focus and decrease the number of mistakes they make. This suggests that social norms can contribute to the strength of a coordinative institution. More specifically, they determine the strength that the institution has over and above that of the social practice.

The significance of social norms can be illustrated by Mackie's (1996) famous study of footbinding in China and how this practice came to an end. Footbinding was introduced at the imperial palace, in all likelihood to promote fidelity control. It spread to the upper class and was copied by the middle and lower classes: 'the higher the social status, the smaller the foot' (Mackie 1996: 1001). Footbinding became popular in spite of its costs: about 10% died of complications; those who survived were pretty much housebound. In fact, bound feet became regarded as a sign of beauty and unbound feet were seen as disgraceful. Only the lowest classes, which needed women to work, formed an exception. Crucially, bound feet were required to be eligible as a wife. The idea was that it fostered fidelity and thereby family honor and it increased paternity confidence. In light of this, Mackie argues that it was a suboptimal coordination equilibrium, which played a central role in the marriage market. The practice prevailed for almost a thousand years. Strikingly, it unravelled within a generation. Mackie takes this to confirm his characterization of the practice as a convention.

Mackie identifies three steps as crucial for its demise: an education campaign, external pressure and the formation of natural foot societies. The education campaign served to raised awareness of just how unhealthy the practice was and

of what the advantages are of unbound feet. External pressure was exerted by other countries who denounced foot binding because of its negative health consequences. The Chinese had adopted all kinds of rationalizations. For instance, women were seen as 'excessively lustful', which is why guarding had become regarded as the natural thing to do. Critical responses to the practice challenged such beliefs and revealed that they were not as compelling as they had come to be seen. Such denunciations questioned the legitimacy of the norm, in particular the normative beliefs that supported it. These challenges to the norm and its sanctions made the existing practice look less attractive, which enabled the Chinese to more fully appreciate the appeal of the alternative. It made them realize, that the prevailing practice constituted a suboptimal equilibrium.

But they did not change the fact that the Chinese were locked into this practice. Given that others participated, the best response remained to do so as well. Parents want their daughters to be successful and that getting married is a substantial part of it. And they jeopardize this by unilaterally deviating from the practice. This problem was ultimately resolved by the formation of natural-foot societies. The members of those societies 'pledged not to bind their daughters' feet nor to let their sons marry women with bound feet' (Mackie 1996: 1011). The underlying idea was: 'the greater the proportion of people who footbind, the less is footbinding a positional advantage in securing marriage' (Mackie 1996: 1012). Natural foot societies provided people with a real alternative due to which violating the norm did not come with the cost of daughters not being able to marry. Those societies turned out to be successful and allowed the new practice of not binding feet to spread. They changed the empirical expectations of their participants (Mackie 1996; see also Bicchieri 2016).

Mackie's account of conventional change confirms that empirical expectations do indeed form the key to the strength of coordination institutions. As it supports SCOOR as a measure thereof. Furthermore, it also reveals how important normative beliefs and sanctions are in this respect. This in turn confirms that *w*/*c* as a second indicator. In addition to this, it illustrates how difficult it is to change a convention. The main problem is that providing information about a better alternative does, as such, nothing to change people's empirical expectations. If people see others continue as before, they have little reason to change their own behaviour.

Even so, it is easy to overestimate the effect of information. Consider 'Chastia', an imaginary society in which people behave in a chaste manner and disapprove of unchaste behaviour (Brennan *et al.* 2013: 25). However, they secretly prefer to behave in unchaste ways, but only if others display unchaste behaviour a well and expect each other to do so. Spiekermann (2015: 177) suggests that 'Chastia could quite easily experience a normative revolution once it became apparent that most Chastians have conditional preferences in line with the lewd norm'. This claim fails to do justice to the lock-in effect of prevailing expectations.

Thus, information is the primary determinant of the strength of coordination institutions, and normativity a secondary one. However, there are others. To the extent that preferences are subject to change, for instance, they affect the strength of a coordination institution as well. Think, for instance, of someone

who already has a few dents in his car. Because of this, he might come to care less about minor accidents and be willing to take more risks.[22] When Lewis' (1969) introduced his account of conventions, he proposed a measure of what he called 'the degree of conventionality', which encompassed all of the elements of his analysis. It consists of six dimensions: the degree of conformity, the fraction of players who expect conformity, the fraction of players who are expected to conform, the fraction of players who prefer to conform, the fraction of players who prefer to conform to the alternative convention, and the fraction of situations in which this is the case (Lewis 1969: 76–80). Each of these factors also bear on the strength of a convention.[23]

In order for any such factor to be an indicator of institutional strength, it has to be compared to a threshold value of that factor, as in SCOOR. Furthermore, I have chosen to single out some of these factors as more important than the others, in particular the extent to which players expect conformity. There is, for instance, little reason to zoom in on the proportion of people who expect conformity, unless some people are in a better position to observe behaviour than others. The final difference concerns the normative dimension of conventions. Lewis (1969: 97) regards conventions as 'a species of norm'. In spite of this, he does not incorporate the effect that norms have on motivation in his measure. I have argued that it is important to do so.

The upshot is that motivation is the main determinant of the strength of cooperation institutions, and information that of coordination institutions. Because of this, the strength of the former can be measured in terms of the weight that participants attribute to a norm, and the latter in terms of the expected degree of conformity.

3. Conclusion

Institutions are norm-governed practices. This view preserves the explanatory power of equilibrium theories. And it captures the normative dimension of institutions in a way similar to rule theories. Crucially, a social norm can govern a social practice even if the latter does not correspond to the former, as long as the participants experience its pull or push. Because of this, the Rules-and-Equilibria Theory can account for weak institutions. As such, it provides a suitable basis for an account of institutional strength. I have argued that this notion should be understood in terms of compliance and robustness. Weak institutions are frequently violated. Effective institutions rarely if ever. And strong institutions are complied with in a robust manner. Because of this, they will remain effective if they meet with substantial challenges.

[22]For this to be the case, he must be tempted to deviate, which means that the situation does not constitute a pure coordination game. In this vein, Sugden (1998) argues that conforming to a convention sometimes runs counter to someone's self-interest.

[23]External shocks can also induce change. Bicchieri (2016) mentions the impact that the contraceptive pill had on the sexual revolution. Consider also the fact that viruses can spread through physical contact. During the coronavirus pandemic, this brought the practice of shaking hands as a way of greeting someone to a grinding halt.

Coordinative institutions are almost always effective if not strong, because the interests of their participants align. In contrast, cooperation institutions are effective or strong only if their norms resolve the conflicts of interests between the participants. Otherwise, they are weak. This difference is due to the fact that coordination institutions are problems of information, whereas cooperation institutions are problems of motivation. Because of this, the strength of these two kinds of institutions should be measured differently. The primary determinant of the strength of coordination institutions is the expected degree of conformity. The main factor that makes cooperation institutions strong is the weight people attribute to its norm. This account of institutional strength sheds light on how valuable institutions can be strengthened and how harmful or unjust institutions can be transformed.

Acknowledgements. I am grateful to Justin Bruner, Francesco Guala and Martin van Hees for helpful comments.

References

Aoki M. 2001. *Toward a Comparative Institutional Analysis*. Cambridge, MA: MIT Press.

Aumann R.J. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* **1**, 67–96.

Becker G. 1968. Crime and punishment: an economic approach. *Journal of Political Economy* **76**, 169–217.

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

Bicchieri C. 2016. *Norms in the Wild*. New York, NY: Oxford University Press.

Binmore K.G. 1994. *Game Theory and the Social Contract: Playing Fair*. Cambridge, MA: MIT Press.

Binmore K. 2008. Do conventions need to be common knowledge? *Topoi* **27**, 17–27.

Binmore K. 2010. Game theory and institutions. *Journal of Comparative Economics* **38**, 245–252.

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

Brennan G., Eriksson L., Goodin R.E. and Southwood N. 2013. *Explaining Norms*. Oxford: Oxford University Press.

Crawford S.E.S. and Ostrom E. 1995. A grammar of institutions. *American Political Science Review* **89**, 582–600.

Cudd A.E. 2006. *Analyzing Oppression*. New York, NY: Oxford University Press.

Ellickson R.C. 1991. *Order Without Law: How Neighbors Settle Disputes*. Cambridge, MA: Harvard University Press.

Fehr E. and S. Gächter 2000a. Fairness and retaliation: the economics of reciprocity. *Journal of Economic Perspectives* **14**, 159–181.

Fehr, E. and S. Gächter 2000b. Cooperation and punishment in public goods experiments. *American Economic Review* **90**, 980–994.

Gintis H. 2007. *The Bounds of Reason*. Princeton, NJ: Princeton University Press.

Greif A. and Kingston C. 2011. Institutions: rules or equilibria? In *Political Economy of Institutions, Democracy and Voting*, 13–43. Berlin: Springer.

Guala F. 2016. *Understanding Institutions*. Princeton, NJ: Princeton University Press.

Guala F. and Hindriks F. 2015. A unified social ontology. *Philosophical Quarterly* **65**, 177–201.

Hart H.L.A. 1961. *The Concept of Law*. Oxford: Clarendon Press.

Haslanger S. 2012. *Resisting Reality*. New York, NY: Oxford University Press.

Hindriks F. 2019. Norms that make a difference: social practices and institutions. *Analyse & Kritik* **41**, 125–146.

Hindriks F. 2021. Rules, equilibria and virtual control: how to explain persistence, resilience and fragility. *Erkenntnis*. https://link.springer.com/article/10.1007/s10670-021-00406-9.

Hindriks F. and Guala F. 2015. 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* **40**, 1–25.
Keuschnigg M. and T. Wolbring 2015. Disorder, social capital and norm violation: three field experiments on the broken window thesis. *Rationality and Society* **27**, 96–126.
Knight J. 1992. *Institutions and Social Conflict*. Cambridge: Cambridge University Press.
Kuran T. 1995. *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Cambridge: Cambridge University Press.
Lewis D.K. 1969. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.
Mackie G. 1996. Ending footbinding and infibulation: a convention account. *American Sociological Review* **61**, 999–1017.
McAdams R.H. 1995. Cooperation and conflict: the economics of group status production and race discrimination. *Harvard Law Review* **108**, 1003–1084.
North D. (1990). *Institutions, Institutional Change and Economic Performance*. Cambridge: Cambridge University Press.
Ostrom E. 2015. *Governing the Commons*. Cambridge: Cambridge University Press.
Pettit P. 1995. The virtual reality of Homo economicus. *Monist* **78**, 308–329.
Pettit P. 2007. Resilience as the explanandum of social theory. In *Political Contingency: Studying the Unexpected, the Accidental and the Unforeseen*, ed. S. Bedi and I. Shapiro, pp. 79–96. New York, NY: New York University Press.
Pettit P. 2015. *The Robust Demands of the Good*. Oxford: Oxford University Press.
Prentice D.A. and Miller D.T. 1993. Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm. *Journal of Personality and Social Psychology* **64**, 243–256.
Rutherford M. 1994. *Institutions in Economics. The Old and the New Institutionalism*. Cambridge: Cambridge University Press.
Sankaran K. 2020. What's new in the new ideology critique? *Philosophical Studies* **177**, 1441–1462.
Schatzki T.R. 1996. *Social Practices*. Cambridge: Cambridge University Press.
Schotter A. 1981. *The Economic Theory of Social Institutions*. Oxford: Oxford University Press.
Shapiro S. 2006. What is the internal point of view? *Fordham Law Review* **75**, 1157–1170.
Spiekermann K. 2015. Explaining Norms (Review). *Economics and Philosophy* **31**, 174–181.
Sugden R. 1986. *The Economics of Rights, Co-operation and Welfare*. Oxford: Blackwell.
Sugden R. 1998. Normative expectations: the simultaneous evolution of institutions and norms. In *Economics, Value, and Organization*, ed. A. Ben-Ner and L. Putterman, 73–100. Cambridge: Cambridge University Press.
Sunstein C.R. 2019. *How Change Happens*. Cambridge, MA: MIT Press.
Tuomela R. 2002. *The Philosophy of Social Practices*. Cambridge: Cambridge University Press.
Ullmann-Margalit E. 1977. *The Emergence of Norms*. Oxford: Clarendon Press.
Young I.M. 1990. *Justice and the Politics of Difference*. Princeton, NJ: Princeton University Press.
Young P.H. 1998. *Individual Strategy and Social Structure*. Princeton, NJ: Princeton University Press.

Frank Hindriks is Professor of Ethics, Social and Political Philosophy at the University of Groningen. He is director of the Centre for Philosophy, Politics and Economics (PPE) and one of the founding editors of the *Journal of Social Ontology*. He has published widely on topics in economic methodology, moral psychology and social ontology. His current research focuses on collective responsibility for sustainable institutions and social justice. URL: https://www.rug.nl/staff/f.a.hindriks/.

Cite this article: Hindriks F. Institutions and their strength. *Economics & Philosophy*. https://doi.org/10.1017/S0266267121000195