Erkenntnis (2023) 88:1367–1389
https://doi.org/10.1007/s10670-021-00406-9

ORIGINAL RESEARCH

Rules, Equilibria and Virtual Control: How to Explain Persistence, Resilience and Fragility

Frank Hindriks[1]

Received: 24 December 2019 / Accepted: 22 March 2021 / Published online: 5 May 2021

Abstract
Institutions are often regarded either as rules or as equilibria sustained by self-interested agents. I ask how these two theories can be combined. According to Philip Pettit's *Virtual Control Theory*, they explain different things: rules explain why regularities persist; self-interest why they are resilient. Thus, his theory reconciles the two theories by adjusting their domains of application. However, the available evidence suggests that rules and self-interest often combine as sources of motivation. Because of this, it is better to integrate the theories rather than to reconcile them. Inspired by Cristina Bicchieri's theory of social norms, I incorporate the notion of rule-following into a game-theoretic account of institutions. According to the resulting *Rules-and-Equilibrium Theory*, institutions are rule- or norm-governed social practices. The theory does not only explain their persistence and resilience, but also their fragility, which provides another reason for preferring the proposed integration to Pettit's reconciliation.

1 Introduction

Institutions, such as languages and traffic rules, involve rules as well as regularities. Furthermore, institutional roles, such as being a police officer or a priest, come with characteristic behaviors that are prescribed by rules. An intuitive way of explaining the regularities is in terms of those rules and the reasons participants have for conforming to them. Strikingly, however, no theory of institutions has fully come to terms with these dual aspects of institutions. Equilibrium theory explains behaviors

This paper is part of a trilogy about norms and institutions. In Hindriks (2019), I introduce the Rules-and-Equilibria Theory, which explains how social norms can motivate in the absence of sanctions. In Hindriks (WiP), I use this theory to explicate the difference between strong and weak institutions. Here I discuss why and how equilibrium theories and rule theories should be combined or unified.

✉ Frank Hindriks
f.a.hindriks@rug.nl

[1] University of Groningen, Oude Boteringestraat 52, 9712 GL Groningen, The Netherlands

in terms of the expectations and preferences of rational agents. It conceives of institutions as behavioral regularities that are due to stable strategies for action.[1] But it does not invoke rules. In contrast, rule theory in general, and theories of rule-following in particular, gives pride of place to the rules that shape how participants conceive of their behavior.[2] But it lacks a general account of why it might be rational for people to conform to them. Hence, the question arises as to how rules and regularities can best be combined into a single theory of institutions.

Thus far, this has been done in two ways. First, the notion of a rule has been integrated into equilibrium theory so as to formulate a hybrid theory (Aoki, 2001; Bicchieri, 2006; Crawford & Ostrom, 1995; Greif & Kingston, 2011; Hindriks & Guala, 2015; Knight, 1992). Most importantly, some of them capture the motivating force that rules can have, even in the absence of sanctions.[3] Second, the two kinds of theories have been reconciled by claiming that they apply under different circumstances. In particular, Pettit (1993), has argued that rule theory explains compliance in ordinary circumstances, whereas equilibrium theory explains it when conformity carries substantial costs. Each of these two strategies has advantages. When the notion of a rule is integrated into equilibrium theory, the resulting hybrid theory offers a precise account of what motivates agents across the entire domain of the theory. However, when the two kinds of theories are reconciled, each is preserved in its entirety. This is particularly valuable because rule theory captures the perspective of the agent or, as Philip Pettit calls it, 'the common mind.' (1995, p. 316) Even so, a reconciliation suffers from an important problem. Its core claim is that in any particular situation, either the one theory or the other explains behavior. Because of this, it cannot do justice to the fact that rules and self-interest often influence behavior in combination.

To resolve this conundrum, I propose a theory that integrates the notion of a rule into equilibrium theory and captures the perspective of the agent, 'the *Rules-and-Equilibrium Theory.*' It has two distinctive features. First, it incorporates the idea that rules as such can motivate. More specifically, starting from Bicchieri's (2006, 2016) theory of social norms, I argue that equilibrium theory can do justice to the idea that people sometimes regard the very existence of a rule as a reason for conforming to it. Second, the *Rules-and-Equilibrium Theory* explains how an institution can consist of both rules and equilibria even when they diverge. It might be, for instance, that people are supposed to bring a bottle of wine to a dinner party. Although people feel the pull of this rule, many are not sufficiently motivated to actually do so. In such a case, the rule of an institution or its sanctions motivates the agents, but not enough to comply. I go on to argue that, even though the two diverge, the rule can still govern the practice.

[1] Equilibrium theories include Ullmann-Margalit (1977), Schotter (1981), Sugden (1986), Young (1998) and Binmore (2010).

[2] Rule theories include Wittgenstein (2010 [1953]), Hart (1961), North (1990), Ostrom (1990), Bloor (1997) and Brennan et al. (2013).

[3] Rules can also play other roles. They can, for instance, help agents interpret their environment (Aoki, 2007; Denzau & North, 1994). Furthermore, they can facilitate coordination by representing signaling devices, such as traffic lights or uniforms, along with suitable responses (Hindriks & Guala, 2015).

I briefly introduce equilibrium and rule theories in Sect. 2. I discuss Pettit's theory in Sect. 3. For reasons I explain there, I refer to it as 'the *Virtual Control Theory*.' In Sect. 5, I compare the two theories and argue that the *Rules-and-Equilibrium Theory* is to be preferred to the *Virtual Control Theory.* This comparison brings their respective advantages and disadvantages into sharp relief and illuminates the significance of the two innovations just mentioned. I conclude that institutions are norm-governed social practices.

2 Theories of Institutions

2.1 Equilibrium Theory

Institutions have two faces: social practices and social norms. Whereas rule theory focuses on the latter, equilibrium theory zooms in on the former. Social practices are regularities in behavior. However, some behavioral regularities are individual rather than social, as when someone habitually cracks his fingers before playing the piano. To be social, the behaviors involved in a regularity must be interdependent, which means that the benefits that one agent incurs, her payoff, depend on what others do (Bicchieri, 2006; Lewis, 1969). Thus, a social practice is a regularity of interdependent behaviors.

Behaviors can be interdependent in different ways. When speaking a language or driving on a particular side of the road, individual interests align. People stand to benefit from coordinating their behaviors with others by doing the same as they do. In other situations, agents face a choice between cooperating and defecting. For instance, sellers on a market often have an incentive to collude and increase their prices so as to make more profit. However, collusions can be fragile because each also faces a temptation to defect (Schotter, 1981; Ullmann-Margalit, 1977). In such cases, their interests or motives are mixed.

The regularities that coordination gives rise to are conventions. Lewis (1969) proposed the first game-theoretic account of conventions. In (pure) coordination games, there are several (equally) good ways of serving the interests of the players, such as driving on the left or the right. A behavioral regularity *R* consists of regular performances of some such action. Such a regularity is relative to a population and a situation that constitutes a coordination game. *R* is a convention exactly if four conditions are met: (1) individuals conform to R, (2) they expect others to conform, (3) they prefer to conform given that others do, and (4) all of this being common knowledge. A convention is a Nash equilibrium, which means that no one will want to deviate from a regularity as long as others do not do so.[4] Performing the relevant action is

[4] In fact, Lewis (1969, p. 14) relies on a stricter solution concept to which he refers as 'a coordination equilibrium.' This entails that, if a player were to deviate, nobody would be better off. Note that Lewis' requirement of common knowledge has been criticized recently and a number of theories no longer include it (Bicchier, 2006; Binmore, 2008, 2010).

rational given that others are expected to do so as well. In other words, preferences for conformity are conditional on expectations of conformity.

The next step in the development of equilibrium theory was to model the normative dimension that conventions often have. This was done by introducing sanctions for failing to conform to a convention. Sanctions can be formal or informal. Think, for instance, of a fine or a frown. They can be modeled by adjusting the payoff of the action for the expected costs of the sanction. In this way, norms can be modeled as costs (Schotter, 1981; Ullmann-Margalit, 1977). As interests converge in coordination games, coordination forms a stable strategy irrespective of a norm. Because of this, all that a coordination norm does is reinforce existing patterns of coordination. In cooperation games, however, interests are mixed. There is a common interest in cooperation, but each individual has reason to defect. Sanctions reduce this tension. In fact, if their expected disutility is high enough, sanctions transform cooperation games into coordination games (Bicchieri, 2006; Ullmann-Margalit, 1977). In this way, they enable cooperation. Irrespective of whether they concern coordination or cooperation problems, equilibrium theory conceptualizes *institutions as interdependent regularities backed by sanctions*.

2.2 Rule Theory

According to rule theory, institutions are rules that constrain behavior. In this vein, Douglas North defines them as 'the humanly devised constraints that shape human interactions' (1990, p. 3). Although some rule theories explain compliance in terms of sanctions, theories of rule-following focus on the reasons that rules themselves provide or on the considerations that justify them. In order to follow a rule, an agent has to conform to it because of these reasons (Bloor, 1997; Brennan et al., 2013; Kripke, 1982; Pettit, 1995, Schatzki, 1996; Wittgenstein, 2010 [1953]). People tend to regard the fact that a rule is in force as a reason for complying with it. On this conception of rule-following, rules cannot be reduced to sanctions, because they come with their own reasons for compliance.

Theories of rule-following address the normative dimension that equilibrium theories ignore. Famously, Wittgenstein (2010 [1953]; cf. Kripke, 1982) attacked dispositional analyses of meaning and mental content arguing that they fail to capture the normativity of rules. In particular, they cannot account for the sense in which behavior that is governed by a rule is correct or incorrect. Another influential argument concerns the legitimacy or authority of legal rules. Hart (1961) criticized the idea that law reduces to the command of the sovereign backed by sanctions. Instead, he proposed to analyze laws in terms of normative rules, which feature rights or obligations. In order to be legitimate, such rules must be accepted by officials. Geoff Brennan, Lina Eriksson, Bob Goodin and Nic Southwood defend a version of this 'Acceptance Theory' for social norms more generally. They maintain that a social norm is 'a normative principle that is generally accepted by the group' (2013, p. 172). In light of this, I conclude that, according to the rule approach, *an institution is a generally accepted normative rule.*

3 The Virtual Control Theory

3.1 A Reconciliation

Theories of rule-following typically explain behavior in terms of what is generally accepted. According to Pettit, they invoke rule-based considerations that are salient to what he calls 'the common mind' (1995, p. 316). Socio-cultural considerations, such as 'this is how we do things here', play a central role in folk explanations: 'we invoke considerations of social acceptance to explain people's abiding by certain norms.' (2007, p. 89) Furthermore, Pettit maintains that the way people deliberate in fact explains what they do: 'The behaviour is actually produced by deliberation over considerations as to what is socially fitting, what is just, or whatever.' (1993, p. 276) Finally, folk explanations pay little attention to self-interest: 'the deliberation [people] practise is not particularly self-regarding in character' (Ibid., p. 272) Thus, rules and rule-following considerations play a central role in how people deliberate and what they do. But self-interest does not. This is why Pettit objects to rational choice theory as an all-encompassing theory of behavior: it represents people as too self-interested.[5]

In spite of this, Pettit does not reject rational choice theory. Instead, he restricts its domain of application. Equilibrium theory represents 'the economic mind', which is predominantly motivated by self-interest (Pettit, 1995, p. 316). However, Pettit proposes, the economic mind is usually only implicitly or virtually present. An agent starts to engage in 'a self-regarding sort of deliberation' only when it is somehow attractive for an agent to deviate from an existing regularity (ibid., p. 321). The idea is that, when socio-cultural factors cease to serve the agent's interests to a sufficient extent, 'the alarm bells of self-interest' ring and conscious decision-making takes over (Pettit, 1995, p. 338). In terms inspired by Simon's (1956) notion of satisficing, the point is that: 'if the course of behaviour adopted by an agent flouts her self-regarding interests, plunging her below a relevant aspiration-level, then she is likely to become aware of the fact and to begin to deliberate in terms of those interests.' (Pettit, 1993, p. 274).

Pettit invokes self-interest to explain why the agent continues to conform to the regularity. At first sight, this is rather surprising. After all, the agent becomes aware of her self-interest exactly because compliance does not seem to serve her interests well enough. However, once she focuses on her self-interest, she will think of other self-interested considerations, such as the disapproval she might face, and they keep her behavior in check. In other words, the common mind or 'the cultural frame' usually explains behavior. Acting on the basis of sociocultural considerations will often have become habitual. But when the risks are too high 'the economic mind' takes over. The agent switches to self-regarding deliberation and acts primarily from

[5] Pettit maintains that, according to rational choice theory, 'people are relatively self-regarding in their desires', which means that: even though they may care about others and what they want, their desires are stronger 'the more [they] bear on their own advantage (1995, p. 311). Thus, when self-regarding motivations are present, selfless motivations play a minor if not negligible role.

desires that bear on her own advantage (Pettit, 1995, p. 310 and p. 321). In this way, the behavior of the agent does not get out of bounds and does not disrupt the existing pattern of behavior. The upshot is that 'the economic mind is *reconciled* with the common mind' (Pettit, 1995, p. 241; emphasis added).

To further clarify how his reconciliation works, Pettit employs an analogy, that of a ball that rolls in a straight line and that is bound to do so because there are side posts along its trajectory.[6] The side posts do not propel it. Instead, they function as standby or virtual causes which ensure that the ball stays within certain bounds and proceeds along its trajectory (Pettit, 1993, 276–277 and 2000, pp. 44–45). Thus, when the common mind is in control, people 'proceed under a more or less automatic cultural pilot in most cases' (Pettit, 2000, p. 240 and 2007, p. 85). However, self-interest is a standby cause that is in virtual control of the regularities. It is active only in exceptional circumstances when a lot is at stake. Thus, people are usually 'virtually self-regarding'; self-regard is then 'in the co-pilot's [seat], ready to assume control.' (Ibid., p. 319 and p. 322) Pettit refers to his account of the economic mind as 'the model of virtual self-regarding control.' (Ibid., p. 319 and p. 320) Because this is its most striking feature, I refer to his overall theory as 'the *Virtual Control Theory*' (*VCT*).[7]

By way of illustration, Pettit discusses slavery. He assumes that slaveowners bought slaves and put them to work because of commonsense considerations. And he observes that slave owners were rarely swayed by moral challenges directed at the practice, even though the arguments used had considerable force. The reason why the institution of slavery usually survived, he goes on to argue, is that slaveowners would realize that discontinuing this practice would cost them dearly (Pettit, 2000, p. 47). The moral challenges are considerations of the common mind. They had a limited effect because, by the time the slaveowners would start to think about giving up their slaves, their economic minds would take over. Thus, slavery persists because it is in line with sociocultural factors. But it is robust to perturbances because of self-interested considerations.

Equilibrium theory is commonly taken to explain both the persistence and resilience of interdependent regularities, both why they are stable and how stable they are (Ullmann-Margalit, 1978).[8] Pettit proposes to restrict its domain of application to resilience.[9] Persistence is to be explained by rule theory. In other words, he adjusts the scope of these theories such that their domains of application no longer overlap. In low-to-moderate risk situations, socio-cultural factors serve the agent's

[6] Another analogue Pettit (2015, pp. 147–49) uses is that of a heating system, which is active only when the temperature in a room deviates substantially from the target level.

[7] Pettit also uses the notion of virtual control to defend particular conceptions of functionalism, democracy, and what he calls 'the robust demands of the good' (1996, 2012, 2015). Here I consider only his use of the metaphor to defend a particular interpretation of rational choice theory.

[8] Ullmann-Margalit criticizes the idea that equilibrium theory explains how interdependent regularities emerge: 'an explanation of this type involves no commitment as to how the scrutinized pattern actually originated.' (1978, p. 284).

[9] Pettit (2007, p. 88) claims that Lewis' (1969) theory of conventions is meant to explain resilience. This interpretation ignores the central role that precedence plays in the theory, which explains persistence.

interests well. In high risk situations, they are unlikely to do so. In this way, Pettit turns the two views from rivals into complements. The following three assumptions reveal how *VCT* reconciles rule and equilibrium theories (*A*1–*A*3):

(*A*1) Rules explain behavior in low-to-moderate risk situations and thereby account for the persistence of social regularities.
(*A*2) Self-interest explains behavior in high risk situations and thereby accounts for the resilience of social regularities.
(*A*3) High risk situations make an agent aware of her self-interest.

These three assumptions have two important implications. The first implication concerns the relation between rules and self-interest insofar as they explain behavior. As *A*1 and *A*2 reveal, the former explain behavior in low-to-moderate risk situations; the latter in high risk situations. This implies (*I*1):

(*I*1) For basically any behavior, there is a single factor that explains it.

I refer to this as 'the winner-takes-all thesis.' The second implication concerns the role of consciousness. According to *A*2, self-interest explains conformity in high-risk situations. *A*3 states that such situations make an agent aware of her self-interest. In other words, the way consciousness works explains how people switch from the common mind to the economic mind. This in turn implies (*I*2)[10]:

(*I*2) When self-interest explains behavior, the agent will most likely be aware of it.

I call this 'the self-interest-awareness thesis.'

3.2 Problems

Both *I*1 and *I*2 turn out to be problematic. A first problem is that, in the examples Pettit uses, self-interest appears to be in actual control all the time. Consider the slavery example. Pettit suggests that, although it made 'economic sense', the institution persisted for non-economic reasons such as 'mere habit', 'a sense of moral commitment', or 'a yen for playing master.' (1993, p. 281) However, slaves provide for cheap labor. Furthermore, people buy slaves on a market where self-interest prevails. To be sure, it may be that slavery is facilitated by cultural factors, but self-interest plays an important role in maintaining the institution (Millar, 2006 [1771], pp. 262–282).

Another example that Pettit (2000, pp. 48–49) discusses concerns someone who reconsiders his membership of a golf club when the fees increase. She doubts that

[10] Rational choice theory as such does not make any assumptions about the role of consciousness in explaining behavior. In fact, it makes few if any assumptions about the psychological mechanisms people rely on when maximizing their utility.

playing golf is sufficiently enjoyable to justify paying the higher fee. But then she realizes that her membership has also been useful for establishing and maintaining professional relationships. Pettit claims that, at this point, the economic mind takes over. However, it appears that, just as in the slavery example, it was in control all along. It is not unusual for people who want to play golf to compare the fees of different golf clubs or of different membership packages. This suggests that the initial decision to become a member was motivated by self-interest, which counts against *A*1.

The winner-takes-all thesis, *I*1, rules out that both rules and self-interest play a substantial role in explaining behavior at the same time. But it seems a real possibility that they do so (Tieffenbach, 2015, pp. 129–30). Bicchieri (2006, 43) argues that, when a norm emerges, people typically conform to it because of sanctions. Over time, however, they come to regard the norm as legitimate. When a norm is well-established, sanctions hardly play a role in explaining conformity. Thus, sanctions and norms can explain conformity together. Their weights change over time. Consider someone who is tempted to cheat but decides not to. He could be motivated merely by the prospect of disapproval in case he was found out. However, he might also refrain from cheating because it is wrong and he already feels guilty just thinking about it. *Pace I*1, it need not be that either of these two factors is sufficient to explain his decision, or even nearly so.

Thus, by pitting social factors against self-regard, Pettit poses a false dichotomy. According to his theory, alarm bells ring when an individual is about to perform an action that goes too much against her self-interest. As a consequence, she switches mindsets. Social factors recede to the background, and self-regard acquires prominence. But how plausible is this? An attractive alternative is to say that both kinds of factors weigh heavily at least some of the time. If so, the winner-takes-all thesis is mistaken.

Turning to the self-interest-awareness thesis, *I*2, Pettit maintains that the common mind usually controls behavior without people being conscious of it. Furthermore, it 'isn't our common experience' that 'human beings [are] rational centres of predominantly self-regarding concern.' (Pettit, 2007, p. 84) However, in those situations where self-interest is the factor that moves us to action, people are usually aware of it. The main rival of rational choice theory, Dual Process Theory, provides reason to doubt this. Many of our decisions are made fast, automatically and without awareness. They are not modulated by deliberation, let alone by moral reasoning. But they are usually self-interested (Evans, 2008; Kahnemann, 2011). The underlying empirical findings reveal that there are strong correlations between low awareness and self-interested motivation. Bicchieri (2006, pp. 4–5 and pp. 47–50) distinguishes between a heuristic and a deliberational route for making decisions in a manner roughly parallel to the two processes of Dual Process Theory. She takes both of them to be consistent with rational choice theory (ibid., p. 47). Furthermore, she claims that 'we often combine the two routes.' (Ibid., p. 5) This supports the claim that, *pace A*3, self-interest often does motivate behavior in low-risk situations. It follows that the self-interest-awareness thesis *I*2 is mistaken.

A further problem from which *VCT* suffers is that, because of Pettit's claim that self-interest explains resilience (*A*2), it remains unclear whether and how the theory

can account for norm-violations. The metaphors he uses—the heating system, the side posts and the autopilot—suggest that it cannot: they all illustrate how something gets back on course. At some point, however, Pettit explicitly allows for the possibility that self-interest disrupts a regularity, when 'the cost of failing to take interests explicitly into account has become too great' (2000, p. 240). The idea that regularities are disrupted by self-interest is rather intuitive. In fact, Tieffenbach (2015, p. 130) claims that, *pace* Pettit, it is part of common sense. However, within the context of *VCT*, it is a surprising claim. According to that theory, people are tempted to deviate from a regularity exactly because they fail to attend to self-interest. This suggests that, once they do, self-interest will make them fall back in line, which is why it explains resilience. Thus, Pettit owes us an account of when and why self-interest causes disruptions.

The upshot is that *VCT* is empirically inadequate. The evidence suggests that both *I*1 and *I*2 are mistaken because *A*1 and *A*2 are false. The main problem is that, even though it combines them in one theoretical framework, *VCT* still treats rules and self-interest as rivals.

4 The Rules-and-Equilibrium Theory

4.1 Social Norms

A number of hybrid theories of institutions incorporate the notion of a rule in equilibrium theory (Aoki, 2001; Crawford & Ostrom, 1995; Greif & Kingston, 2011; Hindriks & Guala, 2015; Knight, 1992). However, none of them captures the normative dimension of behavioral regularities as well as Bicchieri's (2006, 2016) theory of social norms. I use it in Sect. 4.3 to develop a hybrid theory of institutions. But first I introduce Bicchieri's theory in this section. Furthermore, I propose some modifications in Sect. 4.2. Finally, in Sect. 5, I argue that the way in which I integrate rule and equilibrium theories is to be preferred to Pettit's reconciliation.

Bicchieri (2006) is concerned with 'mixed-motives games', such as the Ultimatum Game and the Trust Game. These depict situations where people's preferences are partially coincident and partially opposed. All would benefit from cooperation while it is in the interest of each to defect. Bicchieri's theory of social norms explains why people would cooperate in such situations. She points out that, because behaviors are interdependent, what an agent will do is conditional on what he expects others to do. According to Bicchieri, the object of such 'empirical expectations' is not a regularity but a rule, 'a behavioral rule' that prescribes some course of action (ibid., p. 4). Depending on what he observes, someone might expect others to conform to the rule. In addition to empirical expectations, people can also have *normative expectations*, which are beliefs about what others expect them to do. Someone who possesses a *normative expectation* believes that others expect him to conform to the rule. They will often conceive of this in normative terms, which means that they believe he ought to conform (see Sect. 3.2).

Using these notions, Bicchieri defines the notion of a social norm as follows[11]:

> A behavioral rule R is a social norm in a population P if there exists a sufficiently large subset $P_{cf} \subseteq P$ such that, for each individual $i \in P_{cf}$:
> *Contingency: i* knows that a rule R exists and applies to situations of type S;
> *Conditional preference*: i prefers to conform to R in situations of type S
> on the condition that i possess an empirical expectation and a normative expectation. (Bicchieri, 2006, p. 11)

She specifies the two kinds of expectations as follows[12]:

> (a) *Empirical expectations*: i believes that a sufficiently large subset of P conforms to R in situations of type S; and either
> (b) *Normative expectations*: i believes that a sufficiently large subset of P expects i to conform to R in situations of type S; or
> (b′) *Normative expectations with sanctions*: i believes that a sufficiently large subset of P expects i to conform to R in situations of type S, prefers i to conform, and may sanction behavior. (Ibid.)

Thus, in order for a social norm to exist, people have to know that a rule exists that applies to a particular kind of situation. Furthermore, it must be the case that, if people were to have empirical and normative expectations with respect to a behavioral rule, they would prefer to conform to that rule. Thus, all it takes for a social norm to exist is that people have preferences that are conditional on both kinds of expectations. In light of this, I refer to Bicchieri's proposal as 'the Conditional Preference Theory.'

The next question is what it takes for people to conform to a social norm. In order for this to be the case, they must actually have empirical and normative expectations[13]:

> A social norm R is [practiced] by population P if there exists a sufficiently large subset $P_f \subseteq P_{cf}$ such that, for each individual $i \in P_f$, conditions (a) and either (b) or (b') are met for i and, as a result, i prefers to conform to R in situations of type S.' (Ibid.)

[11] According to Bicchieri, the contingency condition captures the idea that 'collective awareness is constitutive of its very existence as a norm.' (2006, p. 12) The problem with this is that, by postulating such awareness, she seems to assume exactly what needs to be explained. Furthermore, her theory seems to have the resources for doing so, because normative expectations entail such awareness. This suggests to me that it is question begging to include the contingency condition in the analysis. The alternative that I propose below does not include it. Instead, it requires that people have expectations that feature the rule.

[12] For another account of normative expectations, see Sugden (1998). He identifies them with the feelings of approval and disapproval people feel in response to compliance and non-compliance with conventions (ibid., p. 79 and p. 82).

[13] Bicchieri (2006) uses the term 'followed' instead of 'practiced.' But she means nothing more than that the norm is complied with. I use the term 'practiced' to avoid confusion with rule-following proper.

In other words, a social norm is practiced precisely if individuals possess empirical expectations, normative expectations and preferences that are conditional on both of them. Thus, the key idea that the Conditional Preference Theory captures is that a social norm can motivate in the absence of any sanctions.

4.2 The Motivating Power of Social Norms

The Conditional Preference Theory is too permissive: it includes rules that are not social norms. Hence, preferences that are conditional on normative expectations are insufficient for the existence of a social norm. In support of this claim, consider a philosophy department where people could benefit from co-authoring papers. The expertise that faculty members have is complementary in ways that can easily generate synergy among them. Furthermore, they would prefer to cooperate in the following circumstances: they expect others to co-author papers and they believe that others expect them to do so—they might even regard it as obligatory. However, as a matter of fact, people do not conform to the rule and do not believe others expect them to do so. Now, why would this preference structure entail that the rule is a social norm? It may well be that, as a matter of fact, people prefer to work alone, because they abhor the very idea of having to compromise while working on a paper with someone else. It appears that, as things actually are, co-authoring is not a social norm. Instead, it is merely a possible norm.[14]

In an attempt to make the same point, Brennan et al. (2013, pp. 25–26) consider a society they call 'Chastia' where chaste behavior prevails and where unchaste behavior is regarded as inappropriate. In spite of this, the members of this society secretly prefer to live by norms that 'require flagrant displays of unchastity.' (Ibid., p. 25) Brennan et al. go on to argue that Bicchieri's analysis implies that the unchaste rules are norms in Chastia. But they are not. Instead, the chaste rules are norms in Chastia. The example is meant to illustrate that conditional preferences delineate possible social norms, but they do not suffice for actual social norms. The problem with this claim is that, as described, this example concerns an (asymmetric) coordination norm. This entails that it is not a counterexample against the Conditional Preference Theory.[15] I mention it because there is an independent reason to extend the theory from cooperation norms to coordination norms: institutions encompass norms of both kinds. Thus, an account of social norms that is sufficiently general for my purposes should also accommodate examples such as Chastia. This means that the rules

[14] What follows overlaps to some extent with ideas that I present elsewhere (Hindriks, 2019). There are three significant differences between that paper and this one. First, here I develop more explicit and detailed analyses of social norms and institutions. Second, I use them to explain how an institution can be weak or strong. Third, I compare the Rules-and-Equilibria Theory that I defend with Pettit's Virtual Control Theory.

[15] Bicchieri (2006, 38–39) maintains that people need to have empirical expectations in order for a convention to exist. The fact that these are absent in the Chastia example provides for a related reason why it is not a counterexample to the Conditional Preference Theory. Note that, although her definition of social norms is restricted to cooperation norms. Bicchieri (2006, p. 39) argues that conventions can become social norms, for instance when they produce negative externalities (cf. Sugden, 1998).

that feature in the account should not be limited to mixed-motives games but extend to coordination games.

As discussed, the Conditional Preference Theory requires preferences that are conditional on empirical and normative expectations. I propose that, in order for a social norm to exist, people should also have normative expectations. The underlying idea is that, when a social norm exists, people believe that they are supposed to act accordingly. This proposal supports the intuitive verdict that there is no cooperation norm of co-authorship in the philosophy department example just mentioned. After all, the philosophers in that department do not expect others to believe they ought to co-author papers with their colleagues. Furthermore, it also gives the right results with respect to coordination norms, as in the Chastia example. Chastians do not expect others to believe in the unchaste rules. Hence, they do not have unchaste social norms. Now, when normative expectations are widespread, people believe that they are supposed to conform to the rule. But they need not believe that they have reason to comply with it. In other words, they can expect others to believe that they ought to conform without subscribing to the rule themselves. Instead, people who have a normative expectation 'acknowledge' the normative rule.

To make this proposal more precise, I introduce the notion of a normative rule. A normative rule features the notion of an obligation. Its basic structure is $\mathfrak{R}$: 'It is obligatory to do *A* in *S*', or 'Everybody ought to do *A* in *S*.' The notion of a normative rule can be used to explain what a normative belief is. This is the belief that everybody ought to perform a particular type of action in some situation. In other words, it is the belief that $\mathfrak{R}$. This notion can in turn be used to introduce a new conception of a normative expectation. To have a normative expectation is to expect many others to believe that $\mathfrak{R}$.[16] On this conception of them, normative expectations feature obligations, even if only indirectly.[17] The existence of a social norm can now be analyzed as follows [*NE*]:

> [*NE*] A social norm $\mathfrak{R}$ exists in population *P* exactly if a substantial number of members of *P* expect others to believe that $\mathfrak{R}$.

In other words, its normative rule $\mathfrak{R}$ has to be commonly acknowledged. [*NE*] is the first part of what I call 'the Acknowledgment Theory' of social norms.

The Acceptance Theory, which I discussed in Sect. 3.2, insists on normative beliefs. Its core claim is that a social norm is a generally accepted normative rule or

[16] I characterize normative beliefs and expectations as normative because they feature the notion of an obligation, directly or indirectly. They do not as such entail that anyone is obligated.

[17] In this respect, it differs from how Bicchieri defines them. According to Bicchieri's condition (b), a normative expectation is a belief that others have *an empirical expectation*. Note, however, that in her informal discussion of the account she distinguishes two possibilities:

> On the one hand, it might just be an empirical belief. If I have consistently followed R in situations of type S in the past, people may reasonably infer that, ceteris paribus, I will do the same in the future, and that is what I believe. On the other hand, it might be a normative belief: I believe a sufficiently large number of people think that I have an obligation to conform to R in the appropriate circumstances. (Bicchieri 2006, p. 15)

In her more recent work, Bicchieri (2016, p. 35) requires expectations that feature an obligation.

principle (Brennan et al., 2013). The Acknowledgment Theory requires normative expectations instead. Normative beliefs are optional. To see why, consider Oppressia, a society with a totalitarian regime.[18] Oppressians are supposed to express their allegiance to the state whenever an official is present. Although this is an unwritten rule, the penalties for not doing so are fierce. Because of this, the Oppressians conform and expect others to conform. Furthermore, having been brought up in this society, they take it as given that others believe that they ought to do so. But secretly all of them condemn the rule. Proponents of the Acceptance Theory have to deny that the rule of allegiance is a social norm, this in spite of the fact that people are systematically imprisoned for violating it. I take it that intuition is on the side of the Acknowledgement Theory here, which does regard it as a norm. This reveals that normative beliefs are not required.[19]

At the same time, I do think that the notion of a normative belief should play an important role in a theory of social norms. As I have just argued, it is not required for accounting for the existence of a social norm. However, it is needed for understanding how a social norm can motivate in the absence of sanctions. Bicchieri argues that, in order for a social norm as such to motivate someone to comply with it, he has to regard it as legitimate.

And for an individual to perceive a norm as legitimate is for him to take the belief he attributes to others to be 'reasonable', 'legitimate', or 'well founded' (Bicchieri, 2006, p. 21, p. 23 and p. 25). Consider however a meat eater who regards the normative beliefs of vegetarians as reasonable. He can still perceive a norm that requires him to refrain from eating meat as illegitimate. Similarly, someone who practices polyamory may regard monogamous norms as perfectly legitimate, even though she does not subscribe to them. These two examples illustrate that one can regard a normative expectation as justified without perceiving the relevant norm as legitimate. They also reveal what is wrong with Bicchieri's analysis. Perceived legitimacy is not, or at least not only concerned with the beliefs of other people, but also with the beliefs of the agent himself.

Thus, in order for someone to perceive a social norm $\mathfrak{R}$ as legitimate, he has to believe in that norm, by which I mean to say that he has to possess the belief that $\mathfrak{R}$. He subscribes to that norm and believes that he ought to act accordingly. This presupposes that he believes that the norm exists. Given [*NE*], this implies that he also has a normative expectation.[20] Furthermore, in the interdependent situations that social norms are concerned with, an agent has a reason to comply with a normative rule only if others do. Hence, perceived legitimacy also requires empirical

[18] I thank an anonymous referee for this example.

[19] Brennan et al. argue that in some such cases 'the members of the community *mistakenly think* there is a norm' and confuse apparent norms with 'norms proper.' (2013, 35) I disagree. To be sure, those members have *false normative expectations* (see also Bicchieri, 2006, pp. 14–15). However, because of this, they believe that they are supposed to conform to the rule. This suffices for there to be a social norm. One might object that this can be true even if they are not obligated to do so. But this is no different when people have normative beliefs.

[20] Thus, as I conceive of it, someone who accepts a norm also acknowledges it. It is, however, perfectly possible for someone to acknowledge a social norm without accepting it.

expectations. Finally, it must be the case that the agent possesses the normative belief at least in part because of his empirical and normative expectation. To be sure, the normative belief may also be warranted by values, such as safety, productivity, or chastity. But the very fact that a social norm is in force also has to contribute to the justification of the agent's normative belief. Thus, an agent perceives a social norm as legitimate exactly if he possesses a normative belief, an empirical expectation and a normative expectation and he takes the latter two to partly justify the first. Now, these conditions will in principle be satisfied when someone believes in a normative rule, which amounts to believing that it obligates.

In order for a social norm to have motivating force, I propose, an agent has to regard it as legitimate. Furthermore, this fact must have a positive effect on his motivation to comply with it. This requires him to have a normative expectation as well as a normative belief and preferences that are conditional on both. If these conditions are met, the fact that he believes that he ought to comply with the relevant rule motivates him to do so, at least to some extent. This means that the social norm motivates the agent directly.[21] In order for a cooperation norm to be practiced, it has to induce compliance by motivating people directly, indirectly or both. In contrast, conventions are self-enforcing. Their participants are motivated to conform irrespective of any norm that might govern it. However, coordination norms can reinforce an existing equilibrium. They do so when normative expectations make people more motivated to coordinate. In light of this, I propose that for a coordination norm to be practiced, it has to increase their motivation, directly, indirectly or both.

These considerations support a new analysis of the conditions under which a social norm is practiced [*NP*]:

[*NP*] A social norm $\mathfrak{R}$ is practiced within *P* exactly if the members of *P*:
1. expect everyone to conform to $\mathfrak{R}$,
2. expect everyone else to believe that $\mathfrak{R}$,
3. expect that everybody is disposed to sanction violations because of 2, *and/or*
4. believe that $\mathfrak{R}$ partly because of 2,
5. prefer to conform to $\mathfrak{R}$ because of 1 as well 3 and/or 4 [cooperation rule].
5*. prefer to conform to $\mathfrak{R}$ because of 1 and are more favorably inclined to conform to $\mathfrak{R}$ because of 3 and/or 4 [coordination rule].

The three conditions that feature in conditions 5 and 5* capture the motivating role that norms, sanctions and expectations can play.[22] [*NP*] allows for normative beliefs and preferences that are conditional on such beliefs, but does not require them. In

[21] A preference that is conditional on a normative belief will be unconditional only if the agent possesses both a normative and a normative expectation. Earlier I claimed that those normative expectations had to be true (Hindriks, 2019, 141). I thank Guala (2019, 378–80) for pointing out that this is not required.

[22] Thus, apart from norms and sanctions, normative expectations can also motivate. Bicchieri claims that, due to normative expectations, people can 'feel great social pressure' to conform to a rule (2006, p. 14). Presumably, the idea is that the presumed approval and disapproval of others influences them, even if it is not expressed. Such real or imagined effects on how people are esteemed may well increase people's motivation to conform to a social norm.

this respect, the Acknowledgement Theory differs both from the Conditional Preference Theory and the Acceptance Theory. The former does not feature preferences that are conditional on normative beliefs. The latter regards normative beliefs as an essential component of social norms. In contrast, the Acknowledgment Theory allows for people to practice a social norm without subscribing to it.

Finally, the Acknowledgment Theory captures the reason-giving force of social norms better than its rivals. In particular, the idea that preferences can be conditional on normative beliefs can be used to explicate the notion of rule-following, or so I propose. An agent follows a rule exactly if the agent complies to it because of the rule (Brennan et al., 2013). This will be the case when the motivating power of the normative rule is so strong that it induces compliance. Given what I have said about legitimacy, this implies that someone follows a social norm exactly if she conforms to its rule because she regards it as legitimate.[23]

There might be agents who always follow social norms and are never even tempted to violate a rule. I call them 'saints.' Consider also two very different characters: villains and sinners. Villains cannot be bothered to conform to rules. They only do so when sanctions are severe and the probability of being found out is high. In contrast, sinners always feel their pull. But sometimes they succumb to temptation and violate the rules. The Acknowledgement Theory can account for all three characters, because it allows for any combination of the motivating factors discussed. Villains do not believe in normative rules. Sinners do but the extent to which this motivates them is limited. The flexibility of the theory is an important asset of it. In Sect. 5, I argue that, because it is less flexible, *VCT* is not able to account for all three characters, if any.

The key insight that I have defended in this section is that, for a social norm to motivate as such, the fact that an individual perceives it as legitimate must increase his motivation to conform. For him to follow the norm, it must motivate him such that he conforms. In this way, the Acknowledgement Theory of social norms integrates the notion of rule-following view with equilibrium theory.

4.3 Institutions as Norm-Governed Social Practices

Institutions feature social norms. But not every norm is an institution. I propose that, in order for a social norm to constitute an institution, it must affect people's behavior, or at least their motivation. The underlying idea is that institutions can differ in terms of strength. They can be weak or strong. An institution is weak exactly if its norm motivates many of its participants but not enough for them to comply. And it is strong exactly if its norm motivates virtually all of its participants more than enough for them to comply. Thus, full compliance is not required for an institution to exist. Furthermore, it is not sufficient for it to be strong. Imagine that people are supposed to bring a bottle of wine to a dinner party. An institution such as this one is weak when few actually do so. Many might feel the

[23] See Sillari (2012) for an analysis of how Lewis' conception of a convention can be used to explicate other aspects of rule-following.

pull of the norm, but not enough to conform to it. Now, suppose that everybody does so. It will be strong only if people have substantially more motivation than needed for compliance. People will then go out of their way to comply with it. They will bring a bottle even if doing so takes considerable time, effort or money. But it could be that people have just enough motivation to conform. In that case, I will say that the institution is 'effective.' More generally, an institution is effective exactly if it is generally complied with.

As will become important in Sect. 5, differences in institutional strength are closely related to the phenomena that Pettit sets out to explain. Roughly speaking, the regularity that an effective institution generates persists. Furthermore, a strong institution supports a resilient regularity. Such a regularity is very stable and robust to various disturbances, including changes in people's motivation. Pettit does not discuss weak institutions. When an institution is weak, not everybody complies with it. In fact, compliance can be low if not altogether absent. A weak institution might support what might be called 'a gappy regularity', when a noticeable number of people act according to the rule of the institution. Such a regularity is fragile, as it could easily break down completely. Another possibility is that the regularity that is observed does not correspond to the rule at all, as when a particular traffic rule is almost uniformly violated. This regularity will be partly constitutive of a weak institution together with the norm that governs it.

A theory of institutions should be able to explain differences in strength. This entails that its conditions should be stronger than [*NE*] and weaker than [*NP*]. First, a social norm can exist without constituting an institution. This will be the case when it has no effect on people's motivation. Hence, a theory of institutions should be more demanding than [*NE*]. Second, a social norm need not be practiced in order to constitute an institution. Effective and strong institutions are practiced [*NP*]. However, a weak institution is not. All that is required for an institution to exist is that the rule or the sanctions of a social norm provide people with some motivation to comply, even if it is not enough for actually inducing compliance. Hence, a theory of institutions should be less demanding than [*NP*].

In light of these considerations, I propose that an institution is a norm-governed social practice. In order for a social norm to govern a social practice, it must affect the motivation of the participants. However, it need not induce compliance. To make this idea more precise, I assume there is some normative rule $\mathfrak{R}$ that concerns a situation *S*, which constitutes either a coordination game or a mixed-motives game. Given this stipulation, the notion of norm-governance can be explicated as follows [*NG*]:

> [*NG*] A social norm $\mathfrak{R}$ governs a social practice exactly if $\mathfrak{R}$ increases people's motivation to conform to it.

As *S* constitutes a coordination game or a mixed-motives game, the analysis applies to coordination norms as well as cooperation norms. The underlying idea is that, if

people were to comply with the norm, then it would solve a coordination problem or a cooperation problem.[24]

[*NG*] entails that, when a norm governs a practice, people have normative expectations. Without them, a social norm does not influence someone's motivation. In contrast, having a normative belief is optional. When a norm governs a practice, rule $\mathfrak{R}$ makes a substantial number of participants more favorably inclined to conform. This increase in motivation might be due to sanctions, to expectations, or to the norm as such. An interesting example, in this connection, is the speed limit, because it can be violated to different degrees. It might be that people speed on a regular basis, but that they would drive even faster if there were no speed limit at all. In this case, the norm does not only motivate people, but it also affects their behavior. Still, it does not induce compliance, which means that it is a weak institution.[25]

Suppose that $\mathfrak{R}$ is meant to solve a cooperation problem. What does it take for $\mathfrak{R}$ to constitute an institution? One option is that the motivating power of the norm and its sanctions is strong such that people no longer have an incentive to defect. When this happens, people end up complying with the rule. Furthermore, the norm governs the corresponding practice. In this case, the cooperation game has been transformed into a coordination game with multiple equilibria. But what if the motivating power of the norm and its sanctions is less strong and people still have an incentive to defect? The norm can govern a practice even if people defect, which means that the practice does not match the norm or the equilibrium does not correspond to the rule. All that is required is that the norm and its sanctions substantially reduce the incentive to defect or make cooperating substantially more attractive to a non-trivial extent. This is what makes it an institution. For instance, someone who jaywalks might feel the pull of the norm or the push of the sanction. However, these considerations need not be strong enough to prevent him from jaywalking, not even in combination.[26]

This reveals that for an institution to exist, more is required than that people have normative expectations. However, an institution need not be practiced. Thus, the conditions for the existence of an institution are stronger than those for the existence of a social norm [NE] and weaker than those for a norm being practiced [NP]. The following account of institutions meets these conditions [*IE*]:

> [*IE*] An institution $\mathfrak{R}$ exists in a population *P* exactly if a substantial number of its members:

[24] Bicchieri (2006) focuses on cooperation norms and refers to them as 'social norms.' However, she recognizes that conventions can be norms as well (see also Sugden, 1998).

[25] The speed limit is a formal or legal institution. In response to Hindriks (2019), Eriksson (2019) argues that speeding might be an informal norm among those who engage in it. However, the point made in the main text extends to this situation: because the official speed limit motivates, the informal limit will be lower as compared to the situation in which there is no speed limit at all.

[26] An increasing number of norm violations can lead to social change because of how it affects people's empirical expectations. Social change often requires first movers or trendsetters. In this respect, individual differences concerning autonomy, self-efficacy and risk perception play an important role (Bicchieri, 2016, chapter 5).

1. expect a number of others to believe that $\mathfrak{R}$,
2. expect some to do *C* and others to do *D*,
3. believe that $\mathfrak{R}$ partly because of 1, and/or.
4. expect a number of others to be disposed to sanction violations because of 1,
5. are more favorably inclined to conform to $\mathfrak{R}$ because of 2, 3 and/or, and.
6. either doing *C* or doing *D* constitutes an equilibrium in *S*.

This proposal integrates the rule view and the equilibrium view. It entails that institutions involves rules as well as equilibria. But which of them takes priority? Jaywalking might form an equilibrium even though it is prohibited. Should the institution be identified with the rule or with the equilibrium? This question poses a false dichotomy. Both the rule and how people respond to it are part of the institution. The institution is the rule that governs the equilibrium. This holds even if the rule is out of equilibrium. [*IE*] forms the core of what I call 'the Rules-*and*-Equilibrium theory' (*RaE*) of institutions.[27]

Elsewhere Francesco Guala and I have proposed the Rules-*in*-Equilibrium theory (*RiE*; Guala, 2016; Guala & Hindriks, 2015; Hindriks & Guala, 2015). This theory explains coordination in terms of signaling rules. Such rules refer to signaling devices that indicate what to do or whom (not) to approach in the relevant contexts. Think, for instance, of police uniforms, traffic lights and wedding rings. Using '*D*' for signaling devices, their structure is: If S (If *D*, it is obligatory to do *A*). Signaling rules give rise to coordination norms, which govern conventions. As conventions are self-reinforcing, nobody has an incentive to deviate. Because of this, signaling rules are in principle in equilibrium. In this respect, they differ from cooperation norms, which are out of equilibrium when they fail to secure cooperation. *RaE* includes signaling rules. Because of this, it inherits the explanatory power of *RiE*. However, it features an improved account of social norms that does more justice to their normativity. *RiE* models norms as costs: it adjusts the payoffs so as to reflect the effects that norms and sanctions have on the motivation of the agents. In contrast, *RaE* endogenizes these effects by making preferences conditional on expectations about sanctions and beliefs about normative rules. A second difference concerns its scope. It applies not only to coordination institutions but also to cooperation institutions. What is more, it explains how cooperation institutions can be weak, effective and strong.

5 Reconciliation Versus Integration

So, how do the Virtual Control Theory (*VCT*) and the Rules-and-Equilibrium theory (*RaE*) compare? Both explain cooperation in terms of rules as well as equilibria. But they do so in very different ways. I argue that *RaE* is to be preferred to *VCT* for three reasons. First, the way in which *RaE* integrates rule theory with equilibrium theory

[27] Although I do not discuss organized and formal institutions here, *RaE* implies that they are also norm-governed social practices.

is more plausible than how *VCT* reconciles them. Second, *RaE* has more explanatory power than *VCT*, because it explains not only how cooperation institutions can be strong and effective but also how they can be weak. Third, the key implications of *VCT* are false. Because of this, it is empirically inadequate.

To confirm that this third argument speaks in favor of *RaE*, I investigate whether *RaE* has any of the vices that *VCT* has. As discussed, *VCT* has the following two problematic implications:

(*I*1) For basically any behavior, there is a single factor that explains it almost fully.
(*I*2) When self-interest explains behavior, the agent will most likely be aware of it.

RaE is not committed to either of these claims. First, it does not imply that there is a primary motivating factor. Both rules and self-interest, or norms and sanctions, can in principle play substantial explanatory roles in both high and low risk situations. Second, *RaE* is ecumenical with respect to the role of consciousness. In particular, an agent need not be aware of what motivates her behavior, not even when it is self-interest.

Furthermore, *RaE* is inconsistent with two out of the three assumptions of *VCT*. As discussed in Sect. 3.1, these are:

(*A*1) Rules account for the persistence of social regularities, because they explain behavior in low-to-moderate risk situations.
(*A*2) Self-interest accounts for the resilience of social regularities, because it explains behavior in high risk situations.
(*A*3) High risk situations make an agent aware of her self-interest.

RaE accommodates four sources of motivation: (1) expectations, (2) sanctions, (3) norms and (4) self-interested considerations that are independent of (1)–(3). Furthermore, they can be operative at the same time in any combination. It follows that, against *A*1, self-interest can play a substantial role in low risk situations (as when slaves provide for cheap labor). Furthermore, people might attribute so much weight to a norm that it explains compliance even in high risk situations, which is inconsistent with *A*2 (for instance, when they refrain from stealing something valuable when they can get away with it easily). Finally, although *RaE* is consistent with *A*3, it allows for the possibility that self-regard has a significant influence on her behavior when she is not aware of it. The upshot is that *RaE* does not suffer from the empirical problems that afflict *VCT*.

The second reason to prefer *RaE* to *VCT* is that it combines the rule view and the equilibrium view in a more plausible manner. The two theories employ different strategies of unifying the rule view and the equilibrium view. *RaE* integrates them by taking elements from both theories and combining them into a new theory, while rejecting other elements. In contrast, *VCT* reconciles them, which means that each theory is preserved but is assumed to have a distinct domain of application. The idea is, by and large, that persistence can be explained in terms of the rule view, whereas

equilibrium theory explains resilience. For this to work, the explananda must be independent. However, Pettit himself characterizes a resilient pattern as 'a pattern that is robust under various contingencies and that can be relied upon to persist.' (1996, 295) And he points out that 'being in equilibrium, at least for a given context, is a limit case of being resilient.' (Ibid., 297) This reveals that persistence is a matter of stability, while resilience consists of the degree to which a regularity is stable.

It follows that the two explananda are in fact closely related. This creates a problem for explaining them in terms of different factors, which in turn threatens the coherence of *VCT*. Suppose that a rule explains why a particular regularity persists. Now, it might be that it provides more motivation than needed. If it does, it also explains why the regularity is stable to some degree, i.e. why it is resilient. Because this is a real possibility, it is implausible to regard persistence and resilience as distinct phenomena that belong to the domains of different theories. Hence, reconciliation is not a suitable strategy for unifying the rule and equilibrium views. Because it integrates them, *RaE* is more flexible. In principle, any motivating factor can bear on persistence as well as on resilience. Furthermore, the integrated theory allows rules and self-interest to combine as sources of motivation.

The third reason to prefer *RaE* to *VCT* is that it has more explanatory power. This is in fact closely related to the strategies of unification these theories rely on. A successful integration of two theories preserves their key insights and captures new insights by combining some of their elements in original ways. But it discards elements that turn out to be redundant. Because of this, it often explains more in terms of less. Its scope is larger and it is less complex. Thus, it explains phenomena in a more efficient manner as compared to the original theories (Maki, 2001; cf. Friedman, 1974; Kitcher, 1981). In contrast, a reconciliation preserves the integrity of both of the original theories, which means that it is substantially more complex. Furthermore, its scope is the same as that of the two theories combined. Because of this, it is a rather inefficient explanatory strategy.

VCT does indeed combine the explanatory apparatus of the rule view with that of the equilibrium view. Furthermore, it introduces an entirely new assumption (*A*3). In contrast, *RaE* relies on the explanatory apparatus of equilibrium theory. It incorporates the notion of a rule and that of following a rule. But it excludes other factors that feature in rule theories. It follows that *RaE* is less complex than *VCT*. Furthermore, its scope is larger. In particular, it explains not only the persistence and resilience of institutions, but also their weakness or fragility. Thus, it explains the phenomena in a more efficient manner, which means that it has more explanatory power.

The explanatory power of *RaE* can be illustrated further by considering the kinds of characters it can explain. *VCT* is concerned with what might be called 'the prudent person.' This is someone who habitually acts in line with what others do because this is what people are supposed to do. When she is tempted to deviate from a regularity, she comes to realize that this is not in her interest. This explains why she falls back in line. *RaE* also allows for agents to be motivated by norms only in moderate-to-low risk situations. It also accommodates agents who are motivated by sanctions in high-risk situations. Those sanctions can neutralize the force of self-interested considerations that count in favor of violating a norm. Hence, both *VCT*

and *RaE* can account for the prudent person. In this context, the only difference between them is that the agent will still attribute substantial weight to rules when she is tempted to violate them.

However, only *RaE* can account for the behavior of the villain, the sinner and the saint—the three characters that I introduced in Sect. 4.2. According to *A*1, rules explain behavior in low-to-moderate risk situations. But the villain does not regard any norm as legitimate and is not even motivated by them in low-to-moderate risk situations. He violates social norms whenever it suits him. He will refrain from doing so only if violations are frequently detected and heavily sanctioned. According to *A*2, self-interest explains behavior in high risk situations. However, saints follow norms simply because they recognize their force. They are never tempted to violate them. So, no alarm bells ring when their self-interest is at stake. Finally, sinners feel the pull of social norms and they are also sensitive to sanctions. At times, they succumb to temptation and violate a norm when too much is at risk. *VCT* invokes self-interest to explain why people do conform to norms. Furthermore, it does not allow for agents who are conflicted, as sinners are. The underlying problem is that *VCT* hardly allows for individual differences, if at all. Because of this, it is considerably less flexible than *RaE*. The upshot is that *RaE* outperforms *VCT* both with respect to empirical adequacy and explanatory power.[28]

6 Conclusion

Institutions involve both regularities and rules. Equilibrium theories explain under which conditions particular behavioral regularities exist and how stable they are. Rules theories focus more on the normative dimension of institutions, including the role that sanctions play in explaining conformity. Hybrid or unified theories combine both approaches. In this paper, I have compared two such theories: Pettit's Virtual Control Theory (*VCT*) and the Rules-and-Equilibrium Theory (*RaE*). What distinguishes them from other theories is that the rule theories they are concerned with are theories of rule-following, which take rules as such to have motivating power. *RaE* incorporates the notion of rule-following in an equilibrium framework. In contrast, *VCT* combines the theories by taking them to explain different phenomena: rule theory explains the persistence of a regularity, equilibrium theory its resilience.

Although both theories have their attractions, I have argued that the former is to be preferred to the latter. First, *VCT* fails to combine self-interest and rules in a plausible manner. For instance, it assumes that self-interest influences behavior in particular when people are aware of it. However, the available evidence suggests self-interest influences behavior often without people being conscious of this. Furthermore, it mistakenly assumes that, in any situation, one of these is the main motivating factor. In contrast, *RaE* allows for any mix between them. Second, *RaE* has more explanatory power. It is less complex and it explains a wider range of

[28] In Hindriks (WiP), I compare the ways in which *VCT* and *RaE* unify the two theories from a methodological perspective.

phenomena: it also captures the fragility of institutions. Because of this, it does not only account for effective and strong institutions, but also for weak ones.

Acknowledgements I gratefully acknowledge helpful comments from Francesco Guala, Pekka Mäkelä, Philip Pettit, Andreas Schmidt, Emma Tieffenbach and Jack Vromen. I also thank the audiences at the seminar of the Erasmus Institute of Philosophy and Economics (EIPE) in Rotterdam, the workshop on *Social Coordination and Communication* in Nijmegen and the Philosophy of the Social Sciences seminar of the TINT Centre in Helsinki. Finally, I would like to thank four anonymous referees for exceptionally detailed comments that were critical as well as constructive in ways that enabled me to improve the paper substantially.

References

Aoki, M. (2001). *Toward a comparative institutional analysis*. MIT Press.
Aoki, M. (2007). Endogenizing institutions and institutional changes. *Journal of Institutional Economics, 3*(1), 1–31.
Bicchieri, C. (2006). *The grammar of society*. Cambridge University Press.
Bicchieri, C. (2016). *Norms in the wild*. Oxford University 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*. Routledge.
Brennan, G., Eriksson, L., Goodin, R. E., & Southwood, N. (2013). *Explaining norms*. Oxford University Press.
Crawford, S. E. S., & Ostrom, E. (1995). A grammar of institutions. *American Political Science Review, 89*(3), 582–600.
Denzau, A. T., & North, D. C. (1994). Shared mental models: Ideologies and institutions. *Kyklos, 47*(1), 3–31.
Eriksson, L. (2019). Varieties and functions of institutions. *Analyse & Kritik, 41*(2), 383–390.
Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*(1), 255–278.
Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy, 71*(1), 5–19.
Greif, A., & Kingston, C. (2011). Institutions: Rules or equilibria? In N. Schofield & G. Caballero (Eds.), *Political economy of institutions, democracy and voting* (pp. 13–43). Dordrecht: Springer.
Guala, F. (2016). *Understanding institutions*. Princeton University Press.
Guala, F. (2019). Social norms, expectations and sanctions. *Analyse & Kritik, 41*(2), 375–382.
Guala, F., & Hindriks, F. (2015). A unified social ontology. *Philosophical Quarterly, 65*(259), 177–201.
Hart, H. L. A. (1961). *The concept of law*. Clarendon Press.
Hindriks, F. (WiP) Unifying theories of institutions: A critique of Pettit's. *Virtual Control Theory*.
Hindriks, F. (2019). Norms that make a difference: Social practices and institutions. *Analyse & Kritik, 41*(1), 125–146.
Hindriks, F., & Guala, F. (2015). Institutions, rules, and equilibria: A unified theory. *Journal of Institutional Economics, 11*(3), 459–480.
Kahnemann, D. (2011). *Thinking fast and slow*. Farrar.
Kitcher, P. (1981). Explanatory unification. *Philosophy of Science, 48*(4), 507–531.

Knight, J. (1992). *Institutions and social conflict*. Cambridge University Press.
Kripke, S. (1982). *Wittgenstein on rules and private language*. Cambridge (MA): Harvard University Press.
Lewis, D. K. (1969). *Convention: A philosophical study*. Harvard University Press.
Maki, U. (2001). Explanatory unification: Double and doubtful. *Philosophy of the Social Sciences, 31*(4), 488–506.
Millar, J. (2006 [1771]). *The origin of the distinction of ranks*, ed. Aaron Garrett, Indianapolis: Liberty Fund.
North, D. (1990). *Institutions, institutional change and economic performance*. Cambridge University Press.
Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge: Cambridge University Press.
Pettit, P. (1993). *The common mind*. Oxford: Oxford University Press.
Pettit, P. (1995). The virtual reality of homo economicus. *The Monist, 78*(3), 308–329.
Pettit, P. (1996). Functional explanation and virtual selection. *The British Journal for the Philosophy of Science, 47*(2), 291–302.
Pettit, P. (2000). Rational choice, functional selection and empty black boxes. *Journal of Economic Methodology, 7*(1), 33–57.
Pettit, P. (2007). Resilience as the explanandum of social theory. In I. Shapiro & S. Bedi (Eds.), *Contingency.* (pp. 79–96). New York University Press.
Pettit, P. (2012). *On the people's terms*. Cambridge University Press.
Pettit, P. (2015). *The robust demands of the good*. OUP.
Schatzki, T. R. (1996). *Social practices*. Cambridge University Press.
Schotter, A. (1981). *The economic theory of social institutions*. Oxford University Press.
Sillari, G. (2012). Rule-following as coordination: A game-theoretic approach. *Synthese, 190*(5), 871–890.
Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review, 63*(2), 129–138.
Sugden, R. (1986). *The economics of rights, co-operation and welfare*. Blackwell.
Sugden, R. (1998). Normative expectations: The simultaneous evolution of institutions and norms. In A. Ben-Ner & L. Putterman (Eds.), *Economics, value, and organization.* (pp. 73–100). Cambridge University Press.
Tieffenbach, E. (2015). The virtual reality of the invisible hand. *Social Science Information, 55*(1), 115–134.
Ullmann-Margalit, E. (1977). *The emergence of norms*. Clarendon Press.
Ullmann-Margalit, E. (1978). Invisible-hand explanations. *Synthese, 39*(2), 263–291.
Wittgenstein, L. (2010 [1953]). *Philosophical investigations*. Wiley-Blackwell.
Young, P. H. (1998). *Individual strategy and social structure*. Princeton University Press.