Published in *Social Theory and Practice* 48, no. 2 (2022): 285–308.

Making Sense of Full Compliance

Lars J. K. Moen

Abstract

The full compliance assumption has been the focus of much recent criticism of ideal theory. Making this assumption, critics argue, is to ignore the important issue of how to actually make individuals compliant. In this paper, I show why this criticism is misguided by identifying the key role full compliance plays in modelling fairness. But I then redirect the criticism by showing how it becomes appropriate when Rawls and other ideal theorists expect their model of fairness to guide real-world political practice. Attempts to establish institutions conforming to this ideal could have undesirable consequences and might even undermine fairness itself.

Keywords: behavioural economics, compliance, fairness, ideal theory, incentives, principal-agent problems.

1. Introduction

An important assumption in John Rawls's ideal theory is that its principles will attract full compliance across the population. Many critics argue that this assumption makes ideal theory useless, since designing incentives to make people comply with the principles we prescribe is a crucial aspect of political and economic theorizing (Brennan and Pettit 2005; Farrelly 2007; Galston 2010; Levy 2016; Schmidtz 2011). By avoiding this important and challenging issue, these critics argue, ideal theorists can tell us nothing about how institutions ought to operate in the real world, where compliance problems are ever-present.

This paper has two objectives. First, it makes sense of Rawls's full compliance assumption and shows how criticisms of this assumption typically rest on a misunderstanding of its role in ideal theory. Rawls does not assume full compliance to avoid difficult principal-agent problems, as his critics claim. That is, he does not ignore the problems policy makers or regulators (the principal) face when they try to make individuals (the agent) behave as the principles defined in ideal theory require. He instead assumes full compliance in order to stipulate how much each individual must contribute towards realizing a perfectly just society when everyone is equally compliant. Each individual's contribution under such circumstances defines the demands of fairness. Under non-ideal circumstance of partial compliance, fairness cannot require anyone to do more than she or he would need to do to establish or maintain just institutions under ideal circumstances of full compliance. Assuming full compliance is thus a key part in determining what institutions conforming to the ideal of justice as fairness can require of each individual. It is not a cheap way of avoiding hard questions about how to make people compliant.

Rawls actually shows some concern for such questions when he requires that the principles of justice can be reasonably expected to attract full compliance in the society to which they apply. Considering which principles can be reasonably expected to attract full compliance is a part of the process of defining the principles. Rawls and other ideal theorists therefore believe these principles can serve as plausible targets to steer towards in non-ideal theory, which considers how institutions ought to operate under actual conditions (Simmons 2010; Valentini 2009).[1]

The paper's second objective, however, is to reveal practical limitations of the ideal-theory model of fairness and to show why ideal guidance theorists underestimate significant principal-agent problems. On the basis of empirical evidence, I identify two main reasons for why ideal theory cannot guide real-world political practice. First, it will be notoriously difficult to make people behave in the way an ideal principle requires, as they will always find scope for gaming the rules we derive from the principle and for undermining the rules' purpose. Second, attempts to enforce rules of fairness will, under certain circumstances, make everyone worse off, and might even reduce the level of fairness.

As an assessment of the full compliance assumption in ideal theory, the paper thus ultimately shows that while this assumption has a place in a model of fairness, this model cannot guide real-world political practice. Rawls does not assume full compliance simply to avoid principal-agent problems, as his critics claim. However, he does underestimate how such problems restrict the prospect of action-guiding ideal theory. Responding to these problems requires a context-sensitive approach that does not aim towards an abstractly defined target. I thus argue that ideal theory based on the full compliance assumption can constrain non-ideal theory, but it cannot provide a target to steer towards in non-ideal theory. The critics' objection therefore becomes valid if we direct it not at the full compliance assumption but rather at the action-guidance role of the ideal-theory model of fairness.

[1] Holly Lawford-Smith (2010: 360) distinguishes between two ways in which principles can be action-guiding. *Directly* action-guiding principles are supposed to tell us what the right action is all things considered, whereas *indirectly* action-guiding principles are meant to do no more than to contribute in some way to assessing what ought to be done all things considered. As Lawford-Smith notes, Rawls and others taking his approach to ideal theory intend principles to be directly action-guiding. This is particularly evident in Rawls's (1999a: 30–40) rejection of intuitionism, which I discuss in Section 6. In this paper, I therefore take "action-guiding" to mean directly action-guiding. The fairness constraint on non-ideal theory, the plausibility of which I defend in Section 3, can be used in an indirectly action-guiding way, as it can contribute to the consideration of what ought or ought not be done. It cannot, however, determine what should be done all things considered.

This leads me to the view, which I sketch towards the end of the paper, that an ideal-theory model of fairness can be no more than one consideration among others in non-ideal theorizing. We might give fairness special weight, but it must be balanced against other values when we consider how to design social institutions on the basis of how individuals actually respond to incentives under particular circumstances.

2. Full compliance

Rawlsian ideal theory constructs a model of a perfectly just society that is meant to serve at least three purposes in non-ideal theory, in which we make prescriptions for society as it actually is. First, it is supposed to provide an aim to steer towards in non-ideal theory. Without a model of perfect justice as our target, the argument goes, our approach to non-ideal theory will be myopic and we will fail to achieve greater and more distant goods (Simmons 2010). As Rawls (1999b: 90) says, "until the ideal is identified … nonideal theory lacks an objective, an aim."

Second, ideal theory provides a model of perfect justice to which we compare the actual state of affairs. This model is thus meant to help us identify the worst and most pressing cases of injustice and to enable us to see which cases we ought to prioritize. The ideal, Rawls (2001: 13) says, can "help to clarify the goal of reform and to identify which wrongs are more grievous and hence more urgent to correct." The most urgent cases of injustice are "identified by the extent of the deviation from perfect justice" (Rawls 1999a: 216).

Third, as I show in the next section, ideal theory is also meant to formulate a constraint on non-ideal theorizing. The model of fairness developed in ideal theory applies as a constraint on the pursuit of a more desirable society. Specifically, we shall see how ideal theory defines a restriction on what just institutions can demand from individuals. I show in Sections 4 and 5 why ideal theory fails to serve its first two roles. It can, however, perform its third role.

Rawls develops his ideal theory of justice on the basis of several assumptions about the society in which the principles of justice are satisfied. He assumes favourable circumstances under which there is no lack of the "economic means, or education, or the many skills needed to run a democratic regime" necessary for achieving perfect justice (Rawls 2001: 47). He further assumes that society is "a closed system isolated from other societies" (Rawls 1999a: 7) and that people enter society at birth and exit by death (Rawls 2005: 12).

But I shall focus on Rawls's (2001: 13) assumption that "(nearly) everyone strictly complies with … the principles of justice" (see also Rawls 1999a: 7–8). This is also assumed to be common knowledge, so that no one will refrain from contributing due to a lack of assurance that others will take a free ride. Rawls (1999a: 398) further assumes that people have a "sense of justice and desire to do their part in maintaining [just arrangements]." This desire is conditional on the knowledge that others are equally compliant, which gives legal institutions an important role in assuring people that others will do their part (Rawls 1999a: 211). So central is the full compliance assumption in Rawls's ideal theory that he sometimes refers to it as "strict compliance theory" and to non-ideal theory as "partial compliance theory."

This assumption has become a popular target for critics of ideal theory. These critics see no point in an ideal formed on the implausible assumption of full compliance in a world where making people compliant is a central issue we continuously need to deal with. Simply assuming full compliance is a too easy way out of dealing with complicated principal-agent problems. Geoff Brennan and Philip Pettit (2005: 258), for example, argue that assuming full compliance is to assume away the pressing problem of how to design incentives that can actually attract adequate compliance. And David Schmidtz (2011: 778) points out that choosing a principle is to choose a compliance problem. We therefore "cannot set aside compliance as something to address later, because our task of choosing a principle we can live with is a task of choosing a compliance problem we can live with."

In a similar vein, William Galston (2010) understands the full compliance assumption to say that we can move towards the ideal without having to deal with the differences in people's motivation to contribute to the common good. This leads to highly implausible views, he argues, such as the idea that "tax rates can be very high with no consequences for either work effort or tax receipts" (Galston 2010: 405). Colin Farrelly (2007: 845) argues that assuming full compliance is to ignore the problem of determining what is "realistically possible" in a real society where partial compliance is an indisputable fact. And Jacob Levy (2016: 318–319) understands Rawls to assume "away the crime that justifies the state's control of the means of violence, the limited beneficence that sits at the base of theories of justice in property and in the coercive provision of social welfare, and more generally the failings that make politics and justice unavoidable."

We have just seen that ideal theory is supposed to guide non-ideal theorizing about how institutions should actually function in the real world. By assuming away such an undeniably pressing concern in non-ideal theory as partial compliance, ideal theory appears to be a strikingly incapable guide. How can Rawls possibly think otherwise? How can he think an ideal formulated without concern with what will actually attract compliance can nonetheless guide non-ideal theory, where compliance problems are front and centre?

The answer is that he does not ignore these compliance concerns. Rawls does not think we can dream up any utopian ideal and then avoid difficult questions of incentives and motivation by simply assuming full compliance with the ideal. Instead, he says the principles of justice should be defined so that we have good reasons for thinking people can be generally motivated to comply with them. A conception of justice, Rawls (1999a: 398) says, is "seriously defective" unless its principles are defined so that real people can bring themselves to comply with them voluntarily. We therefore need to account for feasibility considerations in ideal theory; we need grounds for assuming full compliance.[2]

This condition is expressed in what Rawls (1999a: 153–160) calls the "strains of commitment," according to which compliance with the principles of justice cannot require citizens to abandon their fundamental interests. It must be in all citizens' personal interest to comply voluntarily, which

[2] Cf. David Estlund (2014: 118): "The fact that people will not live up to [a theory's standards] even though they could is, evidently, a defect of people, not of the theory."

means the institutional structure must protect their ability to pursue their personal ends. Institutional requirements undermining this ability violate the strains of commitment (Rawls 1999a: 153–154). Such institution will make citizens feel alienated from their own society and therefore unmotivated to comply (Rawls 2001: 128). Principles that cannot attract general compliance are therefore not the right principles. As Rawls (1999a: 441) says, if the principles fail to attract full compliance, "some other choice [of principles] might be better."

The rational parties to the original position agree to the principles of justice on the basis of knowledge of facts about economic theory and human psychology that give them an understanding of what people can freely comply with (Rawls 1999a: 119, 137–138). But in spite of this knowledge, they might agree to principles that turn out to be too demanding and therefore incapable of attracting full compliance. Rawls then thinks we must provide the parties with more information and give them another chance to formulate principles with a better chance of attracting full compliance. The principles chosen in ideal theory, or in the original position, may therefore be changed after the veil of ignorance has been lifted and we have turned to non-ideal theory (Rawls 1999a: 105–109, 346, 398, 441, 509). Rawls thus suggests that while ideal theory informs how we ought to do non-ideal theory, discoveries in non-ideal theory also informs how we ought to do ideal theory. Ideal theory receives continuous input from non-ideal theory to make sure that people will be motivated to comply with the principles it formulates. We thus get a feedback loop between the two forms of theorizing (Herzog 2012).

These observations show that Rawls does not assume full compliance as a convenient way of ignoring feasibility concerns, as his critics accuse him of doing. However, this does not explain why Rawls assumes full compliance. Saying that the principles of justice must be capable of attracting full compliance in the actual society to which they apply instead suggests that full compliance is a feasibility constraint on the ideal, not an assumption. Identifying this constraint does not tell us why full compliance is assumed in ideal theory. So, why does Rawls assume full compliance?

3. Fairness

The answer is that the full compliance assumption has a crucial role in Rawls's model of fairness and for determining which institutional measures are permissible in the pursuit of a more desirable society. When just institutions are realized under ideal conditions of full compliance, everyone contributes her or his fair share towards realizing an institutional arrangement from which all benefit. Beneficence beyond such compliance might be desirable, but it is no requirement of fairness and does not make society more just. When we turn to non-ideal theory, where we work with conditions of partial compliance, institutions still cannot demand more from individuals than they would contribute under these ideal conditions.

To get a clearer understanding of how ideal theory formulates this constraint on non-ideal theory, let us first note that the "primary subject of justice," in Rawls's (1999a: 6) view, is "the way in which the major social institutions distribute fundamental rights and duties and determine the division of advantages from social cooperation." These rights protect individuals from being forced to do more than fairness requires, and they have corresponding duties to respect these rights. Fairness requires that individuals behave in accordance with an institutional arrangement insofar as it is just—that is, it promotes common interests expressed in the principles of justice—and insofar as individuals have voluntarily accepted the benefits of this arrangement and taken advantage of the opportunities it provides to further their interests (Rawls 1999a: 96). Those who are compliant and act in this manner do their share in making society a "cooperative venture for mutual advantage" (Rawls 1999a: 4). Rawls (1999a: 301) believes these individuals "have a right to similar acquiescence on the part of those who have benefited from their submission." "We are not to gain from the cooperative labors of others without doing our fair share," he says (Rawls 1999a: 96).

Assuming full compliance is necessary for imagining society as a "cooperative venture for mutual advantage." Under conditions of full compliance, just institutions can function without demanding great sacrifices from particular individuals, since no one is left with an unfairly large share of the cost of establishing and maintaining this cooperative venture. Insofar as everyone benefits from social cooperation, everyone can be expected to contribute her or his fair share in maintaining this enterprise. Under partial compliance, however, the benefits of compliance are smaller, since there will be fewer public goods than under full compliance. To realize the same public goods under partial compliance as under full compliance, some individuals must contribute more than their fair share. Such an arrangement is unfair, as well unstable, since we cannot reasonably expect compliance under such conditions to be in every citizen's personal interests (Rawls 1999a: 296).

Some might, of course, voluntarily choose to contribute more than their fair share to make their society more desirable, but they do not thereby make it more just. Some might, for example, choose to improve the position of the worst off by acts of charity. This is no requirement of justice, however, since justice does not require supererogatory "acts of benevolence and mercy, of heroism and self-sacrifice" (Rawls 1999a: 100).[3] The principles of justice, Rawls (1999a: 495) says, are not "all controlling." They instead leave scope for individuals to pursue their own personal interests at the expense of promoting the overall good of society. A just society, for Rawls, is a venture of fair and equal cooperation, not a maximally desirable society. And since supererogatory acts are not part of a system of reciprocal exchange, they cannot be required by just institutions.[4] A more desirable society, where people do more than their fair share, is therefore no more just than a less desirable society, where citizens only do their fair share.

We can therefore see why consequentialists typically do not engage in ideal theorizing, and how Rawls's theory conflicts with consequentialism. A consequentialist theory can accommodate no such constraint on the pursuit of a more desirable state of affairs—in terms of some agent-neutral

[3] G. A. Cohen (2008) attacks this aspect of Rawls's theory.

[4] We thus see how Rawls derives an agent-centred prerogative from ideal theory that constrains the pursuit of the good. An agent-centred prerogative allows you to systematically assign greater weight to your own interests than to those of others, thus giving you permission to pursue a non-optimal outcome. If the action involves a significant sacrifice to you personally, it might not be the right action for you. Suppose doing *x* would promote the overall good but doing *y* would be more congruent with your interests. You are then permitted to do *y* as long as the personal sacrifice involved in doing *x*, accounting for the extra weight assigned to your own interests, are not outweighed by the loss of the overall good (Scheffler 1992; 1994).

good, such as happiness—as it always favours the more desirable state of affairs. It therefore needs no model of fairness to specify what individuals are required to do, since individuals should always do what best promotes the good, which might plausibly include picking up the slack left by non-compliers. Derek Parfit (1984: 30) is probably right to expect that under non-ideal conditions of partial compliance, we will achieve a better state of affairs if some of us become "pure do-gooders," who are committed to doing what they can to promote the good. For Rawls (1999a: 24), however, requiring such beneficence would conflict with respectful treatment of individuals as separate persons. We have also just seen why such a requirement would conflict with Rawls's emphasis on stability.

The key point here is that contributing what is required for establishing and maintaining just institutions under conditions of full compliance is to contribute one's fair share. The full compliance assumption is thus crucial in modelling the core value of fairness in Rawls's theory. When we then turn to non-ideal theory, where the full compliance assumption does not hold, the notion of a fair share remains the same. No one is required to contribute more towards realizing a just society than would be sufficient under ideal circumstances of full compliance. Demanding more from them would be unfair and therefore no requirement of justice (Murphy 2000: 7).

We thus see how ideal theory constrains the pursuit of a more just society in non-ideal theory. It does so much like Immanuel Kant's (1998: 41–46) Kingdom of Ends. In the Kingdom of Ends, no one is ever treated as a mere means to someone else's ends but always as an end in oneself. Everyone acts on the categorical imperative formulated in the "formula of universal law," which tells us never to act on a maxim we would not want everyone to act on. The perfect moral behaviour of citizens of the Kingdom of Ends models the moral rules that apply also under non-ideal circumstances. Certain actions, such as lying, are famously always wrong, according to Kant, even when they have good consequences.[5] The practice in the Kingdom of Ends thus specifies the moral rules of the non-ideal world.

Christine Korsgaard (1996: 147–151) also recognizes this similarity between Kant and Rawls. They both have "double-level theories," she explains, as they formulate constraints at an ideal level that are imposed onto a non-ideal level. Double-level theories preserve the integrity of persons in the sense that they do not require them to act in undignified ways for the sake of an agent-neutral good, as a consequentialist theory might do. However, Rawls does not permit behaviour in accordance with full compliance without any regard for consequences. Doing so, he says, "would simply be irrational, crazy" (Rawls 1999a: 26). This view is also compatible with at least some of Kant's thought. For example, although he sees the Kingdom of Ends as achievable only under peaceful conditions, he does not think war is never permissible, as that would make nations vulnerable to attacks from their enemies. Kant (1999: 151–158) therefore develops laws of war with perpetual peace as the long-term target.

Building on Korsgaard's analysis of double-level theorizing, Tamar Schapiro (2003) argues that deviations from ideal behaviour are permissible, or indeed required, insofar as they preserve the person's integrity. When others' non-compliance bears on not just the efficiency but also the integrity of compliance, the double-level theory will allow for non-compliance, since compliance will then actually serve an alien will. The conditions of non-compliance undermine the integrity of compliance. Rules that bind categorically are thus nonetheless corruptible and compliance might therefore be a form of complicity under unfortunate non-ideal circumstances. My view of Rawlsian ideal theory takes a similar structure, since institutions cannot require individuals to contribute to a scheme unless it benefits both themselves and others. Only then can they expect their

[5] However, Kant does not, or at least not consistently, think of lying as never justifiable. He says, for example, that "since men are malicious, it is true that we often court danger by punctilious observance of the truth, and hence has arisen the concept of the necessary lie." Stealing, cheating, and even killing, he says, might be necessary in an emergency. And "it rests upon everyone to judge whether he deems it an emergency or not; and since the ground here is not determined, as to where emergency arises, the moral rules are not certain" (Kant 1997: 204). Schapiro (2006) offers a Kantian argument for the permissibility of deception under certain non-ideal circumstances, where your interlocutor has withdrawn her or his commitment to reciprocity.

contributions to be reciprocated. And under any circumstances will requiring them to do more than their fair share compromise their integrity and be incompatible with treating them as separate persons.

Robert Taylor (2009: 488–491) also recognizes that ideal theory is meant to constrain non-ideal theorizing. A consequentialist theory requires agents to perform the actions that best promote an agent-neutral good, but such actions might be irreconcilable with moral intuitions derived from ideal theory, Taylor (2009: 490–491) says. On the basis of these intuitions, Taylor argues that a Rawlsian approach can only support a restricted defence of affirmative action. In particular, ideal-theory constraints might be compatible with monitoring the treatment of certain groups in society, as well as training and mentoring members of disadvantaged groups. However, these constraints will rarely allow for quotas in the pursuit of fair equality of opportunity.

I shall not here consider how ideal theory informs practical treatments of problems under non-ideal circumstances, such as discrimination. My point is rather to complement Korsgaard, Schapiro, and Taylor's constraint views of ideal theory by illuminating how the full compliance assumption is a crucial element in ideal theory as a constraint on non-ideal theory. Specifically, the assumption is crucial for stipulating individuals' fair share in establishing and maintaining a just institutional arrangement and therefore for determining how much they are required to contribute.

We can see, then, how assuming full compliance is not an attempt to conveniently avoid difficult questions of how to make citizens compliant. It is rather an essential step in stipulating what justice as fairness allows institutions to demand from each citizen. Nothing in the common objection to the full compliance assumption introduced in the previous section challenges this reason for assuming full compliance in ideal theory.

4. The problem of indeterminacy

Having clarified the full compliance assumption's role in Rawls's model of fairness, I now turn to the practical limitations of this model.[6] The model gives a rough understanding of what fairness demands of each member of society. All we then need to do is to design rules to make people contribute their fair share. But here we run into problems. We have seen that the critics of the full compliance assumption overlook its role in formulating the constraint on non-ideal theory. And the problems they illuminate do not apply to the constraint role of ideal theory. However, in this section and the next, I show how these are considerable problems for the target role of ideal theory. We have seen that Rawls is aware of problems of motivating compliance, but we shall now see that he is insufficiently alert to these problems. So, while the full compliance assumption is unproblematic as a step in modelling fairness, considerable principal-agent problems emerge when we try to use this model as a guide for non-ideal theory.

I shall focus on two related principal–agent problems. The first is the problem of making rules that are specific enough for making sure individuals behave so as to contribute their fair share towards realizing justice. The second problem, which I discuss in the next section, is that making everyone comply with the demands of fairness might involve a high cost in terms of efficiency, and perhaps even in terms of fairness itself. I identify these problems by considering relevant empirical cases and social-scientific studies. These observations support David Wiens's (2012: 63–64) view that ideal guidance theory makes prescriptions without adequate analysis of causal mechanisms to show how these prescriptions will actually have the intended effects. Indeed, the evidence I find will suggest that no such analysis can adequately support ideal guidance theory.

First, then, there are good reasons for doubting the possibility of gathering all the information necessary for designing rules that can make people actually behave in accordance with rules derived

[6] In this section and the next, I refer to empirical evidence from several cases and studies all of which come from Western societies. While I acknowledge that this cultural limitation prevents me from saying anything universal about human behaviour, I consider it unproblematic for present purposes, since Rawls's focus is mainly on Western liberal, democratic, and pluralistic societies.

from principles defined in ideal theory. A problem is that these rules can generally not specify exactly how agents are to behave. We can expect agents to game the rules—that is, to work around the rules or to find loopholes within them in order to best serve one's own interests—and we typically cannot anticipate how. Institutions guide individuals' behaviour and make some actions more likely than others, but they usually cannot dictate specific actions.

This scope for choice of permissible behaviour might be intentionally built into an institutional structure or discovered by human ingenuity (Goodin 2000). In the latter case, people can undermine the institutions' purpose. The Fourteenth Amendment to the United States Constitution was designed to give emancipated slaves the equal standing and protection as other citizens. In 1886, however, the amendment was used for quite a different purpose in the Supreme Court decision on *Santa Clara County v. Southern Pacific Railroad.* The issue was whether the railroad company owed taxes to the county on certain property, and the Court found in favour of the company. In his defence of the decision, Chief Justice Waite stated that "[t]he Court does not wish to hear argument on the question whether the provision in the Fourteenth Amendment to the Constitution which forbids a state to deny to any person within its jurisdiction the equal protection of the laws applies to these corporations. We are all of opinion that it does" (*Santa Clara County v. Southern R. Co.* 1886). This decision has later been used to protect corporations' right to contribute to political campaigns, a practice affirmed and extended by the Supreme Court in the 2010 case *Citizens United v. Federal Election Commission.*

It is also notoriously difficult to make the language of legal texts specific and rigid enough to avoid interpretation for personal benefit. U. S. Supreme Court Justices often interpret the constitution and statutes differently. Since 1941, the Court's number of unanimous decisions has exceeded 50 percent only four times—in 1996, 1997, 2005, and 2013 (Sunstein 2015: 781).

Problems are likely to persist even if regulators try to make the rules more rigid so as to ensure that they serve their intended purpose. Regulators play an ongoing, and possibly never-ending, game of catch-up with the agents whose behaviour they try to regulate. The regulators will learn from how agents game the rules and adjust the rules responsively, but the agents will then find new ways of gaming the rules, before the regulators again adjust the rules, and so on. Danièle Nouy (2017), then Chair of the Supervisory Board of the European Central Bank, describes such a game of catch-up when she expresses a concern with banks' ability to find new ways of gaming the rules imposed by regulatory bodies. Whatever regulators do to close in on the banks, the banks always find a way to remain out of reach. While in each bank's own interest, such gaming can have devastating long-term consequences for banks collectively and, indeed, for whole economies.

And making rules specific enough to induce the desired behaviour is just one problem. Another problem is to tell who complies the way the regulators want them to and who does not. For example, Christopher Hood (2006; 2007) observes that central government officials in the United Kingdom typically consider it acceptable to deliberately misrepresent their organizations' performances to give the impression that they meet expected standards. Hood further notes that it is very hard to prevent such gaming, as it is difficult to reveal it and even to estimate how widespread it is.

Since institutions typically leave scope for gaming and interpretation for personal benefit, these cases illustrate how individuals will often find ways of complying in ways that benefit themselves, and there is no way of ensuring that it also benefits others. It might be objected, however, that while these cases of gaming show how schemes do not always work as intended, they do not show that the problem of gaming is unavoidable. In Rawls's theory, people are not just rational but also reasonable—that is, they have a sense of justice that informs what ends they rationally pursue.[7] And insofar as this assumption holds, we might think we avoid the gaming problem. In the real world, reasonableness might be less prevalent than in Rawls theory, but perhaps appropriate socialization and education could achieve some progress in this domain.

[7] The two moral powers of reasonableness and rationality are most comprehensively discussed in *Political Liberalism* (Rawls 2005). They are only briefly mentioned in *A Theory of Justice* (Rawls 1999a: 44). However, also in the latter are people assumed to have a sense of justice, which Rawls takes to be the essence of reasonableness.

For this response to work, however, Rawls and other ideal guidance theorists must provide a model showing how these or other measures will have this desired effect. They have so far not done so, and it is far from clear that such a model even can be developed. Of course, we *might* see a move towards less gaming and more reasonableness, but ideal guidance depends on a model that shows why such progress is to be expected. Even if we can show that each actor benefits from a scheme—a "cooperative venture for mutual advantage"—we have not shown that it will not be in the interest of each individual to stretch the rules so as to make the venture most beneficial to themselves, even should this eventually result in the venture's collapse.

And even if every citizen should become reasonable and motivated by a sense of justice, it is still not given that everyone will comply with regulations derived from ideal principles. Rawls is aware that a sense of justice is insufficient for compliance, since reasonable individuals' cooperation is conditional on other individuals being cooperative. Reasonable persons' cooperation for mutual advantage therefore appears to be an assurance game, where everyone is better off contributing but might nonetheless defect insofar as she or he is not assured of others' cooperativeness. Rawls (1999a: 211) therefore stresses the importance of legal institutions assuring people of one another's compliance to ensure the optimal outcome of mutual cooperation. The point illustrated by the cases in this section, however, is that legal institutions cannot provide such assurance. For these institutions to effectively assure agents of each other's compliance, rules must be specific enough about what behaviour they require so that they eliminate the possibility of gaming. We have just seen that it is questionable that they can perform this function, and we therefore have reason to doubt the assurance role Rawls ascribes to these institutions.

These principal-agent problems are feasibility problems, and they might therefore appear solvable for Rawls given his concern with the strains of commitment discussed in Section 2. Complying with the principles of justice cannot conflict with people's personal interests. If we see that the gaming problem I have identified causes an unfair distribution of the burden of maintaining a just society even under full compliance, then we might think we should go back to ideal theory and redefine the ideal so as to make the ideal compatible with the strains of commitment. Accounting for such facts makes ideal guidance seem more realistic. By continuously updating the ideal in response to observations of what we can make citizens voluntarily commit to and not, we appear to make the ideal a more feasible target. As we saw in Section 2, this is how Rawls sees the interplay between ideal and non-ideal theory; he does not leave ideal theory behind for good once he turns to non-ideal theory. Ideal theory is thus continuously updated by input from observations made in non-ideal theorizing (Herzog 2012).

However, relying on this continuous feedback loop between the two forms of theorizing makes the principles defined in ideal theory a moving target it is difficult to steer towards. If an ideal must be continuously redefined in response to empirical facts so as to ensure full, or at least adequate, compliance with its principles, then it cannot perform the long-term guidance role that proponents of ideal theory attribute to it. People will keep discovering new ways of gaming the rules we introduce, which means the ideal cannot be a fixed point in the landscape we can steer towards, but instead a point that continuously changes its location (Rosenberg 2016: 63–65). Making this observation is not to overlook the fact that applying a principle to a real-world case requires assessment and judgment in response to the particular case (Erman and Möller 2013: 28). Rather, the large amount of empirical data that must be taken into consideration when we formulate the principles so as to make them action-guiding will make the theory so context-based it cannot provide a long-term target.

This observation could, in fact, mean the parties to the original position are unable to propose any principles of justice. Rawls (1999a: 125–126) says the parties have knowledge about the "general facts of human psychology" and take this knowledge into account when they agree on the principles of justice in the original position. One such fact is that many people will game the rules derived from these principles. The parties might therefore come to the conclusion that no principle will generally be acted on as they intend, and consequently give up on the task of defining an ideal target for non-ideal theory.

Rawls also admits that an ideal cannot always specify what each of us is required to do. In "the more extreme and tangled instances of non-ideal theory," he says, ideal guidance will "no doubt fail, and indeed, we may be able to find no satisfactory answer at all" (Rawls 1999a: 267). But he does not appear to see this as a major problem for ideal theory. Nor do other proponents of ideal theory. Laura Valentini (2009: 340–341) believes ideal theory can point us in a general direction without dictating everyday decisions, and A. John Simmons (2010: 24) thinks ideal theory's guidance capacity "can reach only as far as our ability to apply it. But that fact constitutes no reason for skepticism about the theory itself."

It is unclear why Rawls, Valentini, and Simmons are so confident that ideal theory can generally provide determinate directives for how to make agents do what fairness requires. Their confidence is not supported by empirical evidence, and we have now seen that such evidence does exist for the opposing view. Here we have therefore revealed how principal-agent problems apply to the target role of ideal theory. We can try to alter the ideal continuously in response to observations of how people are actually motivated to behave, but we then make the ideal a moving target instead of a stable point to aim for in non-ideal theory.

5. The problem of inefficiency

While one problem is to make rules specific enough to avoid gaming, a further problem is that imposing such rigid rules might have undesirable consequences. We might try to incentivize compliance by the use of rewards or punishment, but recent behavioural studies challenge the commonly assumed monotonic relationship between incentives and the level of compliance. I call this the problem of inefficiency, as I show how attempts to enforce rules of fairness could actually be to no one's advantage.[8] It might therefore undermine the ideal of society as a "cooperative venture for mutual advantage" and even the ideal of fairness itself.

[8] By efficiency here, I mean the familiar concept of Pareto efficiency, according to which one feasible alternative, *x*, is more efficient than another, *y*, if and only if at least one person favours *x* to *y* and everyone else either also favours *x* to *y* or is indifferent between the two. In other words, Pareto efficiency requires that someone be made better off if and only if no one be made worse off.

In some cases, however, enforcing fair play is beneficial. This has, for example, proven to be the case in *n*-player public goods experiments (Fehr and Gächter 2000; 2002). In these experiments, the players are typically given $1 and instructed to contribute whatever portion of the dollar they prefer to the common pot. This is all done anonymously. The experimenters then give each player the amount equal to one half of the total amount contributed to the common pot regardless of how much the player contributed. If there are ten players and everyone contributes her or his dollar, then each player receives $5. If one player, A, contributes nothing while all the other players contribute $1, then A, like all the other players, will nonetheless receive $4.50. A will then be better off than the rest, since she ends up with a total of $5.50 and the others $4.50. In fact, the game is designed so that any player will always be better off by contributing nothing no matter what the other players do. The game is repeated in multiple rounds. In these experiments, players typically contribute 50c in early rounds of the game, before their contributions approach zero towards the end of the game, as some players reduce their contributions and other players notice this and respond by reducing their contributions.

The interesting point relevant for present purposes becomes apparent when we compare the results of these experiments to the results of experiments where the players are given the opportunity to punish one another for not contributing to the common pot. This punishment is costly not just to the punished non-contributor, but also to the punisher, who never has a material incentive to punish. Ernst Fehr and Simon Gächter (2000; 2002) call this "altruistic punishment," since it is costly to the punisher but induces potential non-contributors to contribute, thus benefitting everyone else. In experiments with this punishment option, contributions increase throughout the game and approach the maximum of $1 in the final rounds. So, by giving the players the power to punish each other for shirking, fairness is sustained, and everyone is better off than they would have been under conditions of less contribution. In summary, introducing the punishment opportunity causes a sharp and immediate increase in contributions, while removing the punishment opportunity leads to a similarly sharp and immediate decrease in contribution.

But attempts to promote fairness can also be counterproductive.[9] In Boston, the fire commissioner was fed up with fire fighters frequently calling in sick on Mondays and Fridays (Belkin 2002; Greenberger 2003). On December 1, 2001, he responded by cancelling the department's policy of an unlimited number of paid sick days and limiting the number to fifteen per year. Fire fighters exceeding that number would have their salaries docked. The result, surprisingly, was that the number of fire fighters calling in sick on Christmas Day and New Year's Eve increased tenfold compared to the previous year. The commissioner responded by cancelling the fire fighters' holiday bonus checks. The following year, the fire fighters' total number of claimed sick days more than doubled compared to the previous year (from 6,432 to 13,431).

Samuel Bowles (2016) explains this unexpected consequence of punishing non-compliance by referring to the effect of "crowding out." By making the sick-day policy stricter, the fire commissioner made the fire fighters feel disrespected and unappreciated, and therefore made them less motivated to perform their jobs (Bowles 2016: 9–10).[10] But the fire commissioner plausibly acted, at least in part, on a concern for fairness. Some of the fire fighters dutifully showed up for work every day and had to pick up the slack left by those who seemed to take advantage of the fire department's generous sick-days policy. Thus understood, the commissioner tried to enforce fair play. Nonetheless, the fire fighters might have felt unfairly treated, as Bowles suggests, and therefore less motivated to serve their society. If so, their perception of fairness might conflict with the Rawlsian model of fairness, but this should be no surprise, as people in the non-ideal

[9] Ingrid Robeyns (2008: 351) also notes that attempts to promote an ideal might be counterproductive. While a justice-enhancing strategy might look fine on paper, it could make a real-world situation more unjust. For example, Robeyns says, in a hierarchical society with a history of colonial domination, a strategy for enhancing justice might be far more effective if implemented by someone high up in the hierarchy than by someone external to the society. While Robeyns emphasizes that the need for such sensitivity makes ideal guidance complicated, she does not think it makes it hopeless. My argument, however, is that even if we can acquire the information such sensitivity requires, any ideal must be continuously updated in response to this information and therefore cannot be a steady target to steer towards in the long-term.

[10] Bowles (2016) identifies this crowding out effect also in other cases. Brennan and Pettit (2005: 272–274) also discuss crowding out in their critical assessment of ideal theory. However, they do not discuss how it can challenge ideal guidance; they instead focus on how this phenomenon is compatible with their proposed non-ideal theory without ideal theory.

world have conflicting, and often self-serving, understandings of fairness (Babcock and Loewenstein 1997).

We therefore see that even if we can specify in detail what people are required to do, thus solving the problem of the previous section, we might fail to design incentives that will motivate them to act accordingly. In some cases, we will all actually be better off by giving agents some scope for deciding for themselves how much to contribute. While this approach restricts the extent to which we can realize a fair arrangement, it gives agents a sense of control and self-government that has been proven to improve their performance of socially beneficial tasks. Bruno Frey (1997a), for example, shows that levels of tax compliance are high in countries where the authorities allow citizens to declare their own taxes on the assumption that their tax statements are correct. Conversely, levels of tax evasion are high in countries where tax administration is large and costly. By making taxpayers feel trusted, Frey argues, governments can motivate high tax compliance at low monitoring costs.[11]

In his discussion of the Boston Fire Department case, Bowles similarly suggests that the fire commissioner might have been well-advised not to impose stricter penalties but to appeal to the fire fighters' conscience. For example, the commissioner could have simply informed the fire fighters that an inappropriate use of sick days is unfair to those who always show up when they feel well enough to do so. He could thus succeed in creating a motivation of fairness conditional on those in the department who cooperate unconditionally—that is, those who show up for work every day not because of fear of some penalty and regardless of whether others do so or not.

But on the other hand, we should not rule out the possibility that stricter punishment can have a desired effect. Frey (1997b), for example, provides empirical evidence suggesting that while monetary rewards and punishment might crowd out people's moral motivations to behave virtuously, they nonetheless tend to motivate the desired behaviour once they exceed a certain

[11] See Frey (1997b) for more cases where strict law enforcement reduces individuals' "intrinsic motivation" to comply.

size.[12] A problem with this solution, however, is that even if it produces this effect in many cases, it will generally be a costly way for the principal to motivate the agent, as the agent will lack motivation to behave as the principal wants whenever her or his performance is not closely supervised. And in some cases, the required monitoring will not even be possible.

The key to promoting fairness is to figure out which practices will crowd out the motivation to do what fairness requires and under which circumstances, and then avoid such practices. There might always be a way of making citizens behave as the ideal of fairness requires. The problem is that identifying the right incentive structure is complicated and often requires a costly process of trial and error. And we cannot expect ideal theory to tell us whether this cost is worth the benefits, or even if there will be any.

Responding to these practical difficulties calls for a contextual approach sensitive to which attempts to promote fairness are likely to work and under which circumstances. Sometimes, as in the public goods experiments, threatening people with punishment will promote fairness and make everyone better off. In other cases, on the other hand, such as the Boston fire department case, the threat of punishment will lead to more shirking, less fairness, and everyone being made worse off. Rigid rules will sometimes be effective, but at other times, the system should be more flexible.

This suggests a more pragmatic approach to institutional design than Rawlsian ideal guidance. We should not think an ideal defined on the basis of idealized assumptions can point towards a desirable outcome, or towards a "cooperative venture for mutual benefits." While assuming full compliance is helpful for modelling fairness, we should not expect the model to deliver all-things-considered action guidance. We consequently arrive at the same problem as in the previous section: action-guidance requires a more empirically informed approach than is compatible with ideal theory. While we should keep fairness in mind, we should not expect it to always guide how we arrange our institutions.

[12] Uri Gneezy and Aldo Rustichi (2004) identify this effect both in experiments and field studies showing that compared to no reward, small monetary rewards tend to worsen performance. But above a certain threshold, monetary rewards are likely to have the opposite effect.

Rawls therefore faces a dilemma: He can assume full compliance in his model of justice as fairness, but he then cannot plausibly give justice an all-things-considered action-guiding purpose. On the other hand, he can give justice its action-guiding role by making it sensitive to both fairness, efficiency, and other considerations. But then it becomes a target that keeps changing rather than the single shining ideal he thinks we should steer towards. Either way, the full compliance assumption can play a role in modelling fairness. But this model is of limited importance when we consider the appropriate way of structuring social coordination.

6. Values

My point is not that we should forget about abstract ideals and only focus on what is immediately in front of us. We can aspire to making our society more just, more equal, more efficient, and so on, but the rules we prescribe cannot be derived from abstractly defined ideals alone. Rules must be formed by carefully considering how they are likely to affect the behaviour of the agents to whom they apply. Only in these considerations will we come to realize which value to pursue, or perhaps what is the appropriate trade-off between different values. An ideal-theory model of fairness therefore cannot have the practical action-guidance role Rawls attributes to it.

Rawls criticizes "intuitionist theories" that define ideals that cannot tell us what we ought to do, all things considered.[13] These theories give us only a plurality of first principles that might conflict and no priority rule for weighing them when they conflict. Without a priority rule, Rawls (1999a: 30) says, these theories are of "no substantial assistance in reaching a judgment." "An intuitionist conception of justice" does not specify right actions, and is consequently, in Rawls's (1999a: 37) view, "but half a conception". His own ideal theory of justice as fairness, on the other hand, is supposed to provide a basis for determining what the right action is.

[13] With reference to Lawford-Smith's distinction mentioned in fn. 1, we might say that an intuitionist theory is only meant to give us indirectly action-guiding principles, not directly action-guiding principles. These principles identify concerns to be taken into consideration when we decide what to do; they do not specify what ought to be done all things considered.

The last two sections, however, suggest otherwise. Rawls's ideal theory cannot plausibly specify the right action, as we have good reasons for expecting it to fail to successfully inform non-ideal theorizing. Rawls also admits that his theory will not always be determinate. "Precise principles that straightway decide actual cases are clearly out of the question," he says (Rawls 1999a: 319–320). He discusses particularly the issue of civil disobedience and says "a useful theory defines the perspective within which the problem of civil disobedience can be approached; it identifies the relevant considerations and helps us to assign them their correct weights in the more important instances" (Rawls 1999a: 320). But ideal theory cannot assign correct weights to relevant considerations; doing so requires a more contextual and less abstract approach. We have just seen how ideal guidance can lead us to decisions that make everyone worse off.

In his defence of Rawlsian ideal theory and its priority rule based on fairness, Simmons (2010: 28) argues that "so long as it is not clear that such a defense of priority rules must fail—and, in my view, this is not yet clear—we have not been given any strong reason to abandon Rawls's characterizations of either the nature of ideal and nonideal theory or the relationship between them." This is a highly questionable defence of ideal guidance, as it suggests that the burden of evidence is on the critics of this priority rule, not on its proponents. But the preceding discussion has provided evidence against the possibility of ideal guidance. This evidence suggests that while it might be wise to promote fairness in some cases, it is unwise in cases where it will undermine fairness itself as well as the prospect of a "cooperative venture for mutual benefits." The effects on other values should also be taken into consideration. The extent to which we give priority to fairness, or to any other value, should be based on contextual considerations and cannot be determined in ideal theory. The priority rule absent in intuitionist theories is therefore absent also in Rawls's theory.

This view is compatible with applying fairness as a deontic constraint on the pursuit of a more desirable society, as well as with giving more weight to fairness than to other values. But ideal theory cannot tell us when fairness will be outweighed, since that will depend on expected consequences under particular circumstances. The theory can give some priority to fairness, but we need to consult relevant empirical facts to determine when it is outweighed. The facts about compliance I have discussed make it implausible to think we should generally demand what fairness requires.

I should stress here that Rawls himself does not think we should always prioritize fairness. As I noted in Section 3, Rawls (1999a: 26) thinks it would be "irrational, crazy" not to take consequences into consideration.[14] But in light of the preceding discussion, we see that by imposing a sensible restriction on the pursuit of fairness, ideal theory cannot determine when fairness is to be pursued. Ideal theory can only give us a model of fairness; it cannot help us determine how, or to what extent, fairness ought to be promoted.

And a role the ideal-theory model of fairness invariably can perform is as basis for criticizing people for not doing their fair share, even though we might decide against trying to implement rules making them do what fairness requires. We might also use fairness to praise those who do comply. But the ideal-theory model of fairness is no reliable guide in non-ideal theorizing about how institutions ought to operate under conditions of partial compliance.

7. Conclusion

The full compliance assumption is crucial in Rawls's model of fairness, as it is necessary for specifying how much each person has to contribute towards establishing and maintaining ideally just institutions when everyone complies. This level of compliance stipulates a fair share. In non-ideal theory, which deals with partial compliance, justice as fairness requires no one to do more than her or his fair share. Critics typically overlook this purpose of the full compliance assumption in their rejections of ideal theory. They understand Rawls to assume full compliance in order to

[14] Rawls therefore seems to think just institutions can sometimes require that individuals do more than their fair share. Zofia Stemplowska (2016) also takes this view by arguing that justice might require us to take up the slack left by non-compliers. Laura Valentini (2021), on the other hand, argues that justice never requires more than fairness, but we sometimes have a duty of beneficence to take up the slack.

avoid the difficult issue of how to make principles incentive-compatible. We have seen, however, that Rawls assumes full compliance for quite a different reason.

We have also seen that Rawls does take such feasibility considerations into account in ideal theory. If individuals will not be motivated to comply with the principles of justice, we must go back to ideal theory and redefine them. However, here Rawls and other ideal guidance theorists underestimate the force of principal-agent problems. Defining principles we can expect people to actually comply with is notoriously difficult, perhaps even impossible. When we try to use principles defined in ideal theory to guide non-ideal theorizing, we face principal-agent problems akin to those the critics of the full compliance assumption identify. Being responsive to such problems requires ideal theory to be continuously responsive to empirical information to an extent that it loses its long-term guidance capacity. We have seen that people will game the rules derived from principles defined in ideal theory, thus undermining the real-world action-guiding capacity of ideal theory. The solution is to update the theory in response to observations of how people behave. While this is actually what Rawls recommends, he underestimates how much information is needed for making sure the ideal will be action-guiding and capable of attracting full compliance. Accounting for all the relevant information means continuously altering the ideal, thus making it a moving target rather than a fixed point we can steer towards in non-ideal theory.

The second and related part of this objection is that the measures we introduce to promote fairness might be counterproductive. Incentives designed for this purpose could, in fact, lower the overall level of compliance and make everyone worse off. These observations further illuminate the importance of a context-sensitive approach that cannot be guided by an ideal theory. Pursuing fairness, as formulated on the full compliance assumption, might seem desirable when we abstract away from relevant facts about how individuals respond to incentives, but when we account for these facts, we see why we might be better off not trying to steer towards this ideal. The cost of promoting fairness might be intolerably high, even in terms of fairness itself.

These considerations show why Rawls and other proponents of his ideal guidance approach are too ambitious when they expect ideal theory to offer all-things-considered action-guidance. We cannot expect an abstractly defined ideal to guide how we, all things considered, ought to design our institutions. Ideal theory based on ideal assumptions like full compliance can be useful for defining fairness and possibly other values. However, insofar as we want political philosophy to also inform trade-offs between these values, as well as to guide how social institutions actually ought to operate, it must be more empirically informed and more context-sensitive than any ideal theory can be.

School of Politics and International Relations, Australian National University

lars.moen@anu.edu.au

Acknowledgments

I thank Keith Dowding for numerous conversations about ideal theory and Alex Oprea for her insightful comments on an earlier version of this paper.

References

Babcock, Linda and George Loewenstein. 1997. "Explaining Bargaining Impasse: The Role of Self-Serving Biases." *Journal of Economic Perspectives* 11(1): 109–126.

Belkin, Douglas. 2002. "Boston Firefighters Sick—or Tired of Working." *Boston Globe*, January 18.

Bowles, Samuel. 2016. *The Moral Economy: Why Good Incentives Are No Substitute for Good Citizens.* New Haven, CT: Yale University Press.

Brennan, Geoffrey and Philip Pettit. 2005. "The Feasibility Issue." In *The Oxford Handbook of Contemporary Philosophy*, ed. Frank Jackson and Michael Smith. Oxford: Oxford University Press, pp. 258–279.

Cohen, G. A. 2008. *Rescuing Justice and Equality.* Cambridge, MA: Harvard University Press.

Erman, Eva and Niklas Möller. 2013. "Three Failed Charges against Ideal Theory." *Social Theory and Practice* 39(1): 19–44.

Estlund, David. 2014. "Utopophobia." *Philosophy & Public Affairs* 42(2): 113–134.

Farrelly, Colin. 2007. "Justice in Ideal Theory: A Refutation." *Political Studies* 55(4): 844–864.

Fehr, Ernst and Simon Gächter. 2000. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives* 14(3): 159–181.

Fehr, Ernst and Simon Gächter. 2002. "Altruistic Punishment in Humans." *Nature* 415: 137–140.

Frey, Bruno S. 1997a. "A Constitution for Knaves Crowds out Civic Virtues." *The Economic Journal* 107(443): 1043–1053.

Frey, Bruno S. 1997b. *Not Just for the Money: An Economic Theory of Personal Motivation*. Cheltenham: Edward Elgar.

Galston, William A. 2010. "Realism in Political Theory." *European Journal of Political Theory* 9(4): 385–411.

Gneezy, Uri and Aldo Rustichini. 2004. "Incentives, Punishment, and Behavior." In *Advances in Behavioral Economics*, ed. Colin F. Camerer, George Loewenstein, and Matthew Rabin. Princeton, NJ: Princeton University Press, pp. 573–589.

Goodin, Robert E. 2000. "Institutional Gaming." *Governance* 13(4): 523–533.

Greenberger, Scott. 2003. "Sick Day Abuses Focus of Fire Talks." *Boston Globe*, September 17.

Herzog, Lisa. 2012. "Ideal and Non-Ideal Theory and the Problem of Knowledge." *Journal of Applied Philosophy* 29(4): 271–288.

Hood, Christopher. 2006. "Gaming in Targetworld: The Targets Approach to Managing British Public Services." *Public Administration Review* 66(4): 515–521.

Hood, Christopher. 2007. "Public Service Management by Numbers: Why Does It Vary? Where Has It Come from? What Are the Gaps and the Puzzles?" *Public Money and Management* 27(2): 95–102.

Kant, Immanuel. 1997. *Lectures on Ethics*, ed. Peter Heath and J. B. Schneewind. Cambridge: Cambridge University Press.

Kant, Immanuel. 1998. *Groundwork of the Metaphysics of Morals*, ed. Mary Gregor. Cambridge: Cambridge University Press.

Kant, Immanuel. 1999. *Metaphysical Elements of Justice*, second ed., ed. John Ladd. Indianapolis, IN: Hackett.

Korsgaard, Christine M. 1996. *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press.

Lawford-Smith, Holly. 2010. "Ideal Theory: A Reply to Valentini." *Journal of Political Philosophy* 18(3): 357–368.

Levy, Jacob T. 2016. "There Is No Such Thing as Ideal Theory." *Social Philosophy and Policy* 33(1–2): 312–333.

Murphy, Liam B. 2000. *Moral Demands in Nonideal Theory*. Oxford: Oxford University Press.

Nouy, Danièle. 2017. "Gaming the Rules or Ruling the Game? How to Deal with Regulatory Arbitrage." *33rd Société Universitaire Européenne de Recherches Financières Colloquium*, Helsinki, 15 September. URL: https://www.bankingsupervision.europa.eu/press/speeches/date/2017/html/ssm.sp170915.en.html (accessed May 5, 2020).

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

Rawls, John. 1999a. *A Theory of Justice*, revised ed. Cambridge, MA: The Belknap Press of Harvard University Press.

Rawls, John. 1999b. *The Law of Peoples*. Cambridge, MA: Harvard University Press.

Rawls, John. 2001. *Justice as Fairness: A Restatement*, ed. Erin Kelly. Cambridge, MA: The Belknap Press of Harvard University Press.

Rawls, John. 2005. *Political Liberalism*, expanded ed. New York, NY: Columbia University Press.

Robeyns, Ingrid. 2008. "Ideal Theory in Theory and Practice." *Social Theory and Practice* 34(3): 341–362.

Rosenberg, Alexander. 2016. "On the Very Idea of Ideal Theory in Political Philosophy." *Social Philosophy and Policy* 33(1–2): 55–75.

Santa Clara County v. Southern Pacific R. Co. 1886. 118 U. S. 394. *Justitia*. URL: https://supreme.justia.com/cases/federal/us/118/394/ (accessed May 9, 2020).

Schapiro, Tamar. 2003. "Compliance, Complicity, and the Nature of Nonideal Conditions." *Journal of Philosophy* 100(7): 329–355.

Schapiro, Tamar. 2006. "Kantian Rigorism and Mitigating Circumstances." *Ethics* 117(1): 32–57.

Scheffler, Samuel. 1992. "Prerogatives without Restrictions." *Philosophical Perspectives* 6(3): 377–397.

Scheffler, Samuel. 1994. *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions*, revised ed. Oxford: Clarendon Press.

Schmidtz, David. 2011. "Nonideal Theory: What It Is and What It Needs to Be." *Ethics* 121(4): 772–796.

Simmons, A. John. 2010. "Ideal and Nonideal Theory." *Philosophy & Public Affairs* 38(1): 5–36.

Stemplowska, Zofia. 2016. "Doing More Than One's Fair Share." *Critical Review of International Social and Political Philosophy* 19(5): 591–608.

Sunstein, Cass R. 2015. "Unanimity and Disagreement on the Supreme Court." *Cornell Law Review* 100(4): 769–823.

Taylor, Robert S. 2009. "Rawlsian Affirmative Action." *Ethics* 119(3): 476–506.

Valentini, Laura. 2009. "On the Apparent Paradox of Ideal Theory." *Journal of Political Philosophy* 20(3): 332–355.

Valentini, Laura. 2021. "The Natural Duty of Justice in Non-Ideal Circumstances: On the Moral Demands of Institution Building and Reform." *European Journal of Political Theory* 20(1): 45–66.

Wiens, David. 2012. "Prescribing Institutions without Ideal Theory." *Journal of Political Philosophy* 20(1): 45–70.