Political Philosophy

Pevnick, Ryan. 2024. Democratic Stability and Backsliding. *Political Philosophy*, 1(2): 514–544. DOI: https://doi.org/10.16995/pp.17029

Open Library of Humanities

Democratic Stability and Backsliding

Ryan Pevnick, Politics, New York University, US, rpevnick@gmail.com

In light of mounting concerns about democratic backsliding, Rawls's work – which has an unusual focus on considerations of stability – is now being mined for insights about democratic fragility. This article begins by arguing that the key mechanism underlying Rawls's account of stability cannot, consistent with a proper recognition of the burdens of judgment, explain what makes democratic stability possible. It is, therefore, not well-positioned to help us to think productively about how to mitigate the risk of backsliding. Building on an influential literature in political science, I describe an alternative way of thinking about what enables democratic stability that focuses on the importance of giving key actors self-interested reasons for compliance. This account provides a more productive framework for understanding the causes, and thinking about how to mitigate the risk, of backsliding. It should, therefore, contribute to how we evaluate political institutions and proposed reforms.

Democratic Stability and Backsliding

Ryan Pevnick
Politics, New York University, US

Reasonably wealthy democracies were long thought to be nearly invulnerable to breakdown.[1] However, in recent years we have seen "backsliding" even in places where democracy long seemed consolidated.[2] Think of backsliding as the incremental erosion of democratic norms and institutions. Backsliding typically occurs through a series of actions initiated by the executive: making it more difficult for supporters of the opposition to vote, threatening the ability of the media to critically report on the government, eroding the judiciary's independence, and so forth. Notwithstanding its incremental nature, this process may – cumulatively – reduce democratic institutions to a kind of façade adorning an essentially authoritarian regime. Today, backsliding is an important concern facing many democratic countries—Hungary, India, Israel, Turkey, the United States, and beyond.

Understanding backsliding and proposing measures that may mitigate the risk of it ought, then, to be central tasks for democratic theory. Unfortunately, however, most philosophical work in democratic theory ignores considerations of stability and instead focuses almost exclusively on identifying democracy's egalitarian and epistemic virtues. Yet, whatever virtues democracy has on these dimensions will be of little consequence if the system cannot stably persist. As a result, one cannot understand democracy's relative appeal or generate all-things-considered reform proposals without understanding democratic stability and the vulnerabilities associated with backsliding.

In the philosophical literature, John Rawls's work is arguably the most important exception to the general tendency to ignore issues of stability. When pressed to describe the motivation underlying his later work, Rawls explained that he had become increasingly "concerned about the survival, historically, of constitutional democracy."[3] Consequently, *Political Liberalism* is centrally dedicated to exploring "whether in the circumstances of a plurality of reasonable doctrines … a well-ordered and stable

[1] Przeworski et al. 2000, p. 98; Acemoglu and Robinson 2006, p. 56.

[2] Berman 2021, p. 72; Haggard and Kaufman 2021, p. 1.

[3] Rawls 2001a, p. 616.

democratic government is possible."[4] This concern with stability dates back at least to *A Theory of Justice*, which Rawls begins by worrying that, in the absence of a commitment to a shared conception of justice, "distrust and resentment" may "corrode the ties of civility" and "suspicion and hostility" may then "tempt men to act in ways they would otherwise avoid"—thereby undermining the "stability" of democratic regimes.[5]

Recently, some influential figures have turned to the Rawlsian framework to "help in understanding how liberal institutions can reproduce themselves under non-ideal conditions like ours."[6] Given the controversy that surrounds nearly every exercise of Rawls interpretation, some will surely contend that this enterprise misappropriates Rawls's concern with stability, which – one might argue – is instead focused on whether a regime well-ordered by his principles could stably persist. On such a reading, there is a disjunct between Rawls's concern with the stability of *principles* of justice and this article's concern with the stability of democratic *institutions*. This is debatable.[7]

However, my aim is not exegetical. The idea that Rawls's position can illuminate questions of democratic stability is increasingly prominent in the democratic theory literature and so worthy of exploration, regardless of its relationship to Rawls's true views. The article therefore focuses on (a) arguing that the temptation to rely on the Rawlsian framework to help us to understand or address today's pressing concerns of democratic fragility is misguided and then (b) providing an alternative framework that is better suited to these tasks.

The article proceeds as follows. It begins by arguing that the mechanism underlying the Rawlsian account cannot, consistent with due regard for the burdens of judgment, explain how democratic stability is possible. This criticism casts doubt on the account's usefulness in diagnosing the causes of, or helping us to respond to, backsliding (Section

---

[4] Rawls 1993, p. xli.

[5] Rawls 1999, p. 6.

[6] Weithman 2023, abstract; also see Scheffler 2019 and Cohen 2022.

[7] Two points speak against thinking that there is such a disjunct. First, since Rawls's preferred principles of justice require democratic institutions, the two questions are, at least, deeply interwoven. Without a satisfying account of how democracies could stably persist, we cannot hope to explain how a society could, given Rawls's account, be durably just. Second, an interpretation of Rawls that makes him out to be uninterested in the stability of democratic regimes will have difficulty making sense of the pragmatic motivations that Rawls describes in the passages cited above. Indeed, the fact that Rawls talks about stability *both* in terms of the stability of democratic regimes *and* in terms of the stability of principles of justice reinforces the importance of the relationship between the two.

I). The article then turns to, and explains the limitations of, accounts of democratic stability that are instead rooted in a consensus on the value of democratic institutions (Section II). Having criticized these two popular approaches, the article then draws on influential work in political science to sketch an alternative, institutional, explanation of what makes democratic stability possible. This account emphasizes the importance of making democratic institutions "self-enforcing" by giving elite actors self-interested reasons for compliance (Section III). With this account in hand, we can identify strategies that leaders can use, at least under polarized conditions, to erode what would otherwise be the self-enforcing nature of democratic institutions (Section IV). Understanding how democratic stability works, as well as how leaders can effectively undermine it, provides insights into the kinds of responses that can help mitigate the threat of backsliding (Section V). Altogether, then, the article's account allows us to better understand, and think about how to mitigate, backsliding. It also describes a form of stability that, pragmatic foundations notwithstanding, has many of the normatively attractive features promised by the Rawlsian alternative (Section VI).

I. THE RAWLSIAN APPROACH TO STABILITY

This section develops a criticism of the Rawlsian approach to democratic stability. Before doing so, however, it introduces the distinction that Rawls makes between two types of stability and explains the mechanism underlying the Rawlsian account.

A. Two Types of Stability

Stability can be provided by a *modus vivendi*, which exists when the various sides to a conflict understand that further pursuing the conflict will be more costly than any expected gains—perhaps because the sides have similar levels of power. Under such circumstances, it is "wise and prudent" for the sides to avoid open conflict, but they remain "ready to pursue their goals at the expense of the other."[8] Rawls identifies two shortcomings of *modus vivendi* stability. The first is practical: it is not robust to shifts in the underlying distribution of power. The second is normative: in a society stabilized by a *modus vivendi*, citizens are effectively forced to live under institutions that reflect the balance of power in society, rather than institutions and policies that reflect their values.

[8] Rawls 1993, p. 147.

Rawls hopes to show that a democratic society could be stable in a more robust and normatively attractive sense, which he refers to as "stability for the right reasons." Stability for the right reasons exists when it is common knowledge that (a) citizens share a conception of justice and (b) view that conception as a regulative constraint on their behavior. Stability for the right reasons avoids the two weaknesses of *modus vivendi* stability. From a practical perspective, it is superior because citizens "will not withdraw their support of it should the relative strength of their view in society increase and eventually become dominant."[9] From a normative perspective, decisions are not merely "imposed on citizens by brute force";[10] instead, "the reasons from which citizens act include those given by the account of justice they affirm."[11] In this regard, stability for the right reasons makes possible a form of political autonomy that is absent under a *modus vivendi*.

B. Reciprocity and Stability

Having distinguished between these two types of stability, we can now see the sense in which Rawls's account of stability turns on the importance of reciprocity. The difference principle represents a commitment not to "take advantage of contingencies of native endowment, or of initial social position, or of good or bad luck over the course of life, except in ways that benefit everyone, including the least favored."[12] Indeed, Rawls argues that the difference principle is the "only" arrangement, "that meets the following reciprocity condition: those who are better off at any point are not better off to the detriment of those who are worse off at that point."[13] He refers to this as the "deeper idea of reciprocity implicit in" the difference principle.[14] How, though, does the difference principle's reciprocity-based appeal underpin Rawls's account of stability?

To address this question, one needs to consider the sources of instability under the difference principle, as compared to the sources of instability under alternatives, such as restricted utility. Rawls notes that, unlike the difference principle, restricted utility "is a maximizing aggregative principle with no inherent tendency toward either

---

[9] Rawls 1993, p. 148.

[10] Rawls 1993, p. 147.

[11] Rawls 1993, p. xliii.

[12] Rawls 2001b, p. 124; see also Rawls 1999, p. 155–156.

[13] Rawls 2001b, p. 124.

[14] Ibid.

equality or reciprocity."[15] He then argues that because of this lack of a tendency toward reciprocity, there are reasons to worry about whether, under restricted utility, the least advantaged will accept the basic structure. After all, they are asked to have less in order to benefit those who are already better-off for reasons that are arbitrary from a moral point of view (e.g., having been lucky to have been born with socially valued natural aptitude). Rawls says that this can reasonably be expected to lead the least-advantaged to, "grow distant from political society" and to become "withdrawn and cynical."[16] In short, restricted utility's failure with respect to reciprocity poses a threat to stability.[17]

By contrast, in a society organized around the difference principle, it is the *more* advantaged who are "most likely to be discontent."[18] However, Rawls argues that the threat thereby posed to stability is less significant. Such citizens recognize that while they are worse off than they could be, their being so is a natural consequence of a commitment to reciprocity:

> The more advantaged see themselves as already benefited by their fortunate place in the distribution of native endowments, say, and benefited further by a basic structure (affirmed by the less advantaged) that offers them the opportunity to better their situation, provided that they do so in ways that improve the situation of others.[19]

Thus, it is because the difference principle better satisfies the demands of reciprocity than salient alternatives that Rawls takes it to have a stability-based advantage over them.[20]

Importantly, it is not sufficient for this stability argument that the basic structure is, in fact, consistent with a commitment to reciprocity. If the basic structure is organized

---

[15] Rawls 2001b, p. 122.

[16] Rawls 2001b, p. 129.

[17] A different way to put this is to say that restricted utility jeopardizes compliance because, on Rawls's account of moral motivation, citizens are willing to abide by *fair* terms of cooperation when they are proposed. But, in view of its violation of reciprocity, restricted utility organizes the basic structure so as to further benefit those who have already been advantaged with respect to the contingencies of natural talent, family background, or luck over the course of their lives (Rawls 1993, 81).

[18] Rawls 2001b, p. 125.

[19] Rawls 2001b, p. 126.

[20] In keeping with this perspective, Samuel Scheffler argues that the United States now has "growing resentment and discord, and the degrading destabilizing of liberal institutions" because it "has egregiously failed to live up to any reasonable standard of reciprocity" (Scheffler 2019, 4).

in accordance with the difference principle, but the well-off *think* that that principle imposes an unfair burden on them (perhaps because they disregard the contingent nature of their advantages), then it will not be conducive to stability. Thus, although it is not often recognized, it is critically important from the perspective of the Rawlsian stability argument that citizens *perceive* the basic structure to be consistent with a commitment to reciprocity and respond to it accordingly.

This is important because if there is disagreement about justice, then it is virtually assured that there will also be disagreement over whether or not the basic structure is consistent with a commitment to reciprocity. (After all, many of those who deny that the difference principle is called for by justice are also likely to see it as inconsistent with the demands of reciprocity.) A shared conception of justice is, therefore, central to the Rawlsian account of stability because it allows for a shared perception that a society's basic structure satisfies the demands of reciprocity. And, this shared perception is, on Rawls's account, critical for stability.

C. A Dilemma for the Rawlsian Approach to Stability

This section will argue that the Rawlsian account faces an important dilemma that undermines its ability to explain democratic stability. If we interpret the overlapping consensus as requiring convergence on a particular conception of justice, it is – in light of the burdens of judgment – infeasible. Yet, if we interpret the consensus broadly enough to satisfy worries related to feasibility, then the reciprocity-based mechanism that the Rawlsian account of stability relies upon should no longer be expected to operate. Simply put: *the mechanism underlying the Rawlsian account of stability is not compatible with due regard for the burdens of judgment*.

Begin, then, with the narrower conception of the overlapping consensus, which requires convergence on a particular conception of justice. If everybody accepts justice as fairness, then considerations related to reciprocity can stabilize the society by explaining to both the least- and most-advantaged why they ought to comply. This possibility coheres with Rawls's description of the advantage that justice as fairness has over alternatives with respect to stability.

However, such an account faces an important feasibility-based challenge. Over time, Rawls came to view the expectation – endorsed in Part III of *A Theory of Justice* – that all citizens may come to accept the same comprehensive doctrine as "unrealistic" and "utopian" because it conflicts with the burdens of judgment.[21] These burdens include

[21] Rawls 1993, p. 39.

the difficulty of assessing and weighing conflicting and complex considerations, the inevitable vagueness of moral and political concepts, the way in which our personal experiences shape our perspective, and the wide variety of relevant considerations. Yet, these burdens also affect people's reasoning about questions of justice. Thus, as has often been noted, just as it is unrealistic (even under favorable conditions) to expect convergence on a given comprehensive doctrine, so too it is unrealistic to expect convergence on a given conception of justice.[22] The implication is that even if a shared conception of justice among citizens could help underwrite stability for the right reasons by leading all citizens to view the basic structure as satisfying the requirements of reciprocity, this would require a uniformity of views that is unrealizable in democratic political life.

In light of the challenges associated with achieving consensus on a *particular* conception of justice, it is tempting to interpret the consensus as merely requiring agreement on a *family* of conceptions of justice that share certain features. This could include the enumeration of certain rights, the assignment of some priority to those rights, and a commitment to ensuring that all citizens have adequate means to make use of such rights. Rawls embraces such a modification in his later work, asserting that it is "more realistic" to suppose that the focus of an overlapping consensus would be "a class of liberal conceptions."[23] This approach may be buttressed by the observation that although citizens did not share a commitment to a particular conception of justice even during democracy's most stable periods, those periods may seem to have been underwritten by a narrowing of the range of disagreement about justice.

This approach plausibly allows a society that is divided, for instance, between followers of Rawls and followers of Milton Friedman to count as having an overlapping consensus. Friedman can reasonably be interpreted as arguing for the strong protection of certain basic liberties and for a negative income tax that provides all citizens with a floor level of resources.[24] If, as an example, (nearly) everyone in the society accepted one of these conceptions of justice, then there would be agreement on some priority for the basic liberties and the need for citizens to have adequate means to make use of them, even while there remained considerable disagreement about the best interpretation

[22] For arguments along these lines see, for example, Klosko 1993, p. 350; Waldron 1999, p. 152; Thrasher and Vallier 2018, p. 399; and Cohen 2022, p. 18.

[23] Rawls 1993, p. 164.

[24] Friedman 1991, pp. 191–194. While there is room for disagreement about how best to interpret Friedman's position, Friedman is just standing in here for a kind of restricted utility position that is likely to arise in a democratic society. Nothing ultimately hangs on whether the historical Friedman, in fact, endorsed such a position.

of these commitments. Because positions that insist on protecting a floor level of resources for all, while embracing utilitarianism above that floor, should be expected to attract support in a free democratic society, a plausible account of democratic stability needs to be compatible with their presence. By recognizing this, the broader conception of the overlapping consensus may overcome the feasibility-based concerns faced by the narrower conception.

Unfortunately, the broader conception purchases this realism at the cost of undermining the power of the reciprocity-based mechanism on which the Rawlsian account relies. To see this, notice that Friedman's position – which may win out in the democratic politics of such a society – suffers from the stability-based liabilities that Rawls associates with restricted utility and its failure to satisfy the demands of reciprocity. Under such a system, Rawls says, the least advantaged "cannot affirm the principles of justice" in their "thought and conduct over a complete life."[25] This is because, as noted above, they see their fellow citizens supporting, and benefiting from, a basic structure that gives the less advantaged lower life expectations than they would have had under available alternatives, and does so for reasons that ought to be considered arbitrary from a moral point of view (e.g., their being born to a less advantaged position). As we have seen, Rawls argues that a society that organizes itself around restricted utility therefore risks the least advantaged growing "distant from political society" and becoming "withdrawn and cynical."[26] The concern is that restricted utility attempts to maximize social welfare without taking seriously enough the cost that some are asked to bear for the sake of others and, as a result, risks alienating those citizens. The critical point is that interpreting the consensus broadly enough to escape concerns of feasibility comes at the cost of undermining the reciprocity-based mechanism for stability that Rawls's account depends upon.

Here, then, is the dilemma—either Rawlsian accounts require too much uniformity of belief to plausibly arise in a large democratic society or they admit positions that are inconsistent with the reciprocity-based mechanism for stability upon which they rely. Might there be a "goldilocks solution" to this dilemma—that is, a consensus that allowed for the reasonable disagreement about justice that is sure to arise in a democratic society, even while excluding views that raise concerns of instability by failing to satisfy the demands of reciprocity? The discussion of restricted utility should make one skeptical. Bracketing its ultimate appeal, restricted utility seems like a relatively obvious candidate conception of justice—the kind of possibility that is likely

[25] Rawls 2001, p. 128.

[26] *Ibid*.

to arise and attract support in any democratic society that allows for free expression and vibrant political disagreement. If, as Rawls argues, even *that* possibility is inconsistent with the reciprocity-based mechanism of stability, then it seems doubtful that one can rely on that mechanism to stabilize a democratic society while also being realistic about the extent of disagreement that is likely to arise.[27]

The broad implication of the discussion in this section is that the Rawlsian explanation is not well-positioned to help us to understand what could facilitate stability in real democratic societies. Yet, if the Rawlsian account is not well-positioned to help us to understand what makes democratic stability possible, it is also unlikely to help us understand, or think productively about, how to respond to democratic breakdown.

II. CONSTITUTIONAL CONSENSUS

In light of the difficulties associated with pinning our hopes for democratic stability on Rawls's reciprocity-based mechanism, it seems worth exploring the possibility, instead, of stability built around ordinary citizens' endorsement of democratic institutions. In an early article on Rawls's "political" turn, Kurt Baier argued that while actual democracies do not feature consensus on a particular conception of justice, they do often achieve consensus "on the procedures for making and interpreting law."[28] Similarly, George Klosko has argued that survey evidence in the United States shows little agreement on maximizing the position of the least well-off, but (now dated) evidence of "diffuse" support for democratic regimes as a whole.[29] The central thought animating these positions is that democracies will be stable if ordinary citizens come to a consensus on the value of democratic institutions because citizens will view themselves as having strong reasons to comply with democratically-enacted laws—even when, substantively, they disagree with those laws. The idea that a broader endorsement of

[27] A different way to put this point would be to say that regardless of whether we classify Friedman's position as belonging to the family of liberal political conceptions that could constitute an overlapping consensus, the Rawlsian account faces an important problem. If Friedman's position counts as a part of an overlapping consensus, then people who are part of the consensus will disagree about whether the society's institutions satisfy the reciprocity constraint. The consensus, then, will be too broad to make the underlying stability mechanism operable. Meanwhile, if restricted utility views of the kind held by Friedman do not count as part of the consensus, then the account of stability is incompatible with a realistic assessment of the diversity of views of justice that is likely to arise in such a society.

[28] Baier 1989, p. 775; cf. Barry 1995, p. 910.

[29] Klosko 1993, p. 356.

democratic institutions can contribute to stabilizing democracy has a long history in political science[30] and is supported by some empirical work.[31]

The relative feasibility of a constitutional consensus notwithstanding, two concerns about this approach limit its usefulness from the perspective of understanding and mitigating the risk of backsliding. First, existing accounts are atheoretical in the sense that they do not describe the mechanism by which a constitutional consensus can emerge and be sustained over time. Clearly, such a consensus will not *always* exist—many societies are deeply divided about the appropriate procedures for selecting leaders and making laws. Without understanding what can enable such a consensus, we cannot say anything about the kinds of steps that a society should take to facilitate democratic stability or prevent backsliding.

The second shortcoming of such approaches is that they do not seem to recognize that important mechanisms that drive backsliding can operate *even in the presence* of a constitutional consensus. It is critical to recognize this, since, in many cases, backsliding occurs despite citizens overwhelmingly supporting democratic institutions. Indeed, a crucial strategy that leaders use to erode democratic institutions – call it *stealth*[32] – involves their undermining democratic institutions without citizens being aware that they are doing so. For instance, a government may take steps to make it more difficult to vote in neighborhoods in which opposition supporters tend to live, without these steps being widely recognized. The government may also openly take actions that undermine the democratic operation of the system as a whole, while creating uncertainty about the implications of those actions. For instance, the incumbent may make it difficult for opposition supporters to vote by adopting regulations that are ostensibly necessary to prevent voter fraud. Even where stealth does not succeed in *fully* hiding undemocratic actions, creating a degree of opacity makes it easier for the government's supporters to rationalize the relevant actions as compatible with democracy.[33] Since a constitutional consensus cannot be counted on to disable the strategies associated with stealth, it is a mistake to think that such a consensus is, by itself, sufficient to explain democratic stability or to robustly protect against backsliding.[34]

---

[30] Lipset 1959; Easton 1965.

[31] Claasen 2020. The idea is also closely related to Habermas's (2001) influential idea of Constitutional Patriotism, though I cannot engage with the large literature surrounding Habermas's position here.

[32] Luo and Przeworski 2023.

[33] Krishnarajan 2023.

[34] One might object that insofar as a leader works to undermine democratic institutions, this shows that there is not, in fact, a constitutional consensus: by taking such actions, the leader

The implication is not that a constitutional consensus is of *no* value when it comes to explaining democratic stability. To the contrary, we will return to the idea of a constitutional consensus in Section V.A, where I will argue that an appealing account of democratic stability will stress the complementary aspects of accounts rooted in constitutional consensus and accounts rooted in proper institutional design. Thus, my claim in this section is only that focusing on the importance of a constitutional consensus is *insufficient* as an explanation of democratic stability and, therefore, as a framework for conceptualizing problems related to backsliding.

III. INSTITUTIONS, INCENTIVES, AND STABILITY

Building on influential work in political science, this section describes how certain types of institutions – elections, enumerated rights, and an independent judiciary – can play an important role in stabilizing a political system by rendering it such that even self-interested agents have powerful reasons for compliance.[35] Not only can such an account go a long way towards explaining how stable democratic government is possible, it can also help explain how a constitutional consensus might arise. This is important, as we shall see, because while the stability provided by well-designed institutions may – especially under polarized conditions – be vulnerable to enterprising politicians, a constitutional consensus can make it more difficult for such figures to erode the relevant institutions. Thus, the article's account of stability emphasizes the mutually reinforcing role of well-designed institutions and a commitment, amongst ordinary citizens, to those institutions.

---

reveals themselves not to share in the relevant commitments. Yet, if we understand a constitutional consensus to require that *any* individual who comes to power will be so intrinsically committed to democratic institutions that they will not attempt to undermine them, even when they can gain enormously by doing so, then a constitutional consensus is – in practice – likely an unrealizable goal. It effectively depends on people becoming invulnerable to the corrupting effects of power. If the idea of a constitutional consensus is interpreted in a way that is effectively unrealizable, it will be of little help in thinking about how actual democratic societies might mitigate the threat of backsliding. Accordingly, I choose to think of a constitutional consensus as holding when the vast majority of ordinary citizens share the relevant commitments. This is an interpretation of the idea that seems – to me, at least – more consistent with realistic agency assumptions.

[35] I do not wish to claim that this is the only constellation of institutional arrangements that can have such a stabilizing effect, but it is a historically important possibility (*cf.* Levy 2007; Kogelmann 2017; Ober 2017, ch. 2; and Carugati 2020).

A. Elections

I begin by explaining the crucial role that elections can play in rendering a political system stable. In a sense, it is puzzling that incumbents hold elections and that losing candidates comply with the results. While some may do so because they are intrinsically committed to the fairness of electoral procedures, it is inevitable that others will lack a sufficiently strong normative commitment of that kind. If electoral systems are to determine who controls government, there must be reasons to hold elections and comply with the results that *even these individuals* find compelling.

Such individuals will compete electorally when they view their expected return from doing so as superior to the return that they will get from attempting to subvert the system.[36] But, why – from this perspective – would competing factions view electoral competition favorably?

Begin with the opposition. They can organize to subvert the existing system so that they do not need to accept its results. However, doing so is likely to be dangerous and to carry with it remote prospects for victory. Thus, if the government is willing to hold reasonably fair elections, the opposition will often find competing in them attractive.[37] These elections must be reasonably fair to attract opposition participation—not because the opposition is assumed to be independently wedded to norms of electoral fairness, but because the opportunity to participate in rigged elections cannot substitute for the appeal of regime subversion.

Why, though, would the government offer its opposition the opportunity to compete on reasonably fair terms? The simplest case is that the government is very confident that it will win elections, but – by holding them – it induces the opposition to compete by the rules of the game rather than work to subvert the regime extra-institutionally. The government may prefer this because – from its perspective – attempts at regime subversion carry some (even if low) probability of a cataclysmic outcome—perhaps involving the execution or expulsion of its leading officials. Furthermore, even when they fail, attempts to subvert the regime may be costly to the government in other

[36] See, influentially, Przeworski 1991 and Acemoglu and Robinson 2006.

[37] The important exception is when opposition elites (a) perceive the choice of regime type to be high stakes and (b) expect to reliably lose elections. As Acemoglu and Robinson (2006, ch. 7) argue, these are the conditions under which opposition elites may seek to partner with the military in attempted coups. A simplifying assumption in their framework, that the identity and interests of elites is exogenous to political institutions, may lead to understating the stabilizing power of electoral systems.

ways—for instance, they may destabilize the economy, leading capital (human and financial) to flee.

Why, though, might the governing party hold elections and comply with the results even when it believes it is likely to lose? It may do so when (A) the costs associated with the opposition's attempts to subvert the regime outweigh the expected benefits of canceling elections and (B) holding elections can be expected to forestall efforts at regime subversion. (A) is more likely to obtain when the governing party anticipates that (1) it will have a realistic prospect of competing for, and gaining access to, office in the foreseeable future[38] and (2) its important interests will be protected even if it surrenders office. In sum, canceling elections – even when one expects to lose – may not be particularly appealing if doing so: carries with it a significant risk of revolt and severe accompanying personal costs, promises to destabilize the regime in ways that will undermine one's economic position,[39] and when one's rights – including the right to compete to regain office – will anyway be protected.

Notice, furthermore, that when the government openly works to subvert an existing electoral system (e.g., by censoring journalists, stripping opposition supporters of the right to vote, engaging in electoral fraud, etc.), this provides evidence that it lacks the public support to win election outright. After all, the incumbent who expects to handily win a fair election has reason to *avoid* fraud and the questions that it will raise about the extent of their actual public support. Thus, engaging in such behavior provides a public signal that the incumbent lacks broad support, which can simultaneously raise doubts about their competence and help coordinate opposition on non-electoral efforts at removal.[40]

Finally, an electoral system makes it more difficult for a leader who wishes to subvert the existing regime to get support from other elites within their coalition for doing so. In a typical non-electoral system, such elites will depend on the leader for their power. It will, as a result, be difficult for these elites to resist the leader. By contrast, in electoral systems, many other elites will depend for their power, at least directly, on the support of their own constituents.[41] This gives them a freer hand to resist their own party leaders

---

[38] Przeworski 1991.

[39] This consideration may help explain the remarkable historical stability of wealthy democracies—these are cases in which political elites have a lot to lose, economically, from regime instability.

[40] Fearon 2011.

[41] The extent to which this is true varies across electoral systems and depends, more broadly,

than they would have in a typical non-electoral system. Moreover, they will often have reason to resist such attempts because they have valuable positions, and in many cases, reasonable hope for competing for higher office, in the existing system. Were the existing regime effectively subverted, not only is there is no reason for confidence that they would be as well-placed within its replacement, but they would also need to worry that the leader would regard them – given their popularity – as potentially dangerous competitors, the kind who new autocrats often work to eliminate. Thus, the existence of elections makes it more difficult for party leaders to get elite support, even from members of their own coalition, for attempts at regime subversion.[42]

This subsection explains why opposition and governing elites will often prefer to compete within an electoral system, as well as why – once such a system is in place – it becomes particularly difficult to subvert. For our purposes, the crucial point is that political elites will often have very strong self-interested reasons to comply with electoral institutions. These reasons may often be dispositive even in the *absence* of a moral consensus on: a particular comprehensive doctrine, a conception of justice, or even the intrinsic fairness of the electoral system itself. While an influential stream of political science literature[43] has emphasized the potentially self-enforcing nature of electoral systems, this remains – in the philosophical literature – an oft-overlooked virtue of democratic systems, one that competing regime types (e.g., lottocracy, epistocracy, etc.) cannot be presumed to share.

To be sure, however, while we can identify mechanisms associated with elections that facilitate regime stability, the presence of elections – as is evident from any number of nation-building exercises – is not *sufficient* for such stability. The section's more modest claim is that elections can stabilize a regime by incentivizing key players to focus on competition within the existing rules of the game. The next section describes supplementary institutions that can, when properly designed, enhance the stabilizing virtues of electoral systems.

---

on the political environment. For instance, the nationalization of politics in the United States has increased the dependence of local officials on national party leadership, which has made local officials a less reliable check on the anti-democratic actions of leadership (Pierson and Schickler 2024, p. 168).

[42] This paragraph casts doubt on the suggestion (e.g., Waldner and Lust 2018, p. 100) that institutions are epiphenomenal, with democratic stability depending *simply* on the background distribution of power between factions.

[43] E.g., Przeworski 1991; Weingast 1997; Fearon 2011.

B. Enumerated Individual Rights and an Independent Judiciary

Elections will have difficulty securing participation if contestants view the consequences of defeat as unacceptable. Individual rights can help protect against this possibility by assuring those who lose elections that their most fundamental interests are nevertheless secure. However, since it cannot be assumed that everybody who gains office will willingly respect such rights, the commitment to protecting them must itself be made credible. Is there an explanation, then, of why a government would avoid violating such rights that does not depend on officeholders being intrinsically committed to protecting them?

Enumerated rights and an independent judiciary can make it costly for the government to violate individual rights. To see why, following Weingast,[44] imagine a setting in which the government benefits from retaining office and from violating citizens' rights (think of this as successful rent-seeking). The citizens, who are divided into two factions, do best when their rights are not violated. They can, if they jointly resist the government, overthrow it—thereby avoiding the costs associated with rights violations and leaving the government with its worst outcome. Resistance, however, is costly (albeit less costly than suffering rights violations). Under these specifications, citizens face a coordination problem with two pure strategy equilibria. If one faction knows that the other faction will acquiesce, then their best response is to also acquiesce (unilateral resistance would be costly, but unsuccessful). In such a setting, the government's best response is to violate rights. However, if either faction will reliably challenge a government that violates rights, then the other's best response is to do so as well, since their resistance will lead to the government being overthrown. In such circumstances, the government's best response is to respect rights.

From this perspective, an obstacle to realizing the equilibrium in which the government respects rights is disagreement between the factions over what constitutes a rights violation. To avoid this, citizens need a shared understanding of what constitutes a rights violation and whether such a violation has occurred. In the absence of a robust shared moral view, enumerated rights can serve as a focal point, helping citizens to coordinate on what they will "count" as a rights violation for the purpose of coordinating resistance.[45] Meanwhile, the judiciary can provide an independent signal,

---

[44] Weingast 1997.

[45] For this to work, the enumerated rights must – at least, broadly – correspond to rights that people, in fact, value—otherwise citizens will have no reason to exercise costly resistance in response to their violation. An implication is that in circumstances in which factions lack

indicating whether or not the government has violated enumerated rights. Together, then, enumerated rights and an independent judiciary can help to create a shared understanding among citizens about whether the government has violated rights, which can help citizens overcome their coordination problem. Since the government can anticipate this effect and wishes to retain office, it is incentivized to avoid behavior that would bring forth such a ruling. This contributes to democratic stability directly because some such rights are necessary preconditions of democratic government and, indirectly, by lowering the stakes of electoral loss.

Having now sketched this institutional account of democratic stability, it is important to emphasize that there is no reason that a Rawlsian cannot accept it. After all, the Rawlsian is anyway committed to the importance of elections, enumerated rights, and an independent judiciary—albeit on different, and presumably independently sufficient, grounds. The point, then, is not that the institutional account requires one to take steps that the Rawlsian would reject, but – rather – that it draws attention to a set of mechanisms that are ignored by the Rawlsian explanation of stability, and that can be expected to be operative even in an environment marked by the diversity of conceptions of justice that one should expect in a democratic society. Unlike the Rawlsian account, then, the institutional alternative is well-positioned to help us to understand democratic stability and, because of that, to see how backsliding works and how its likelihood might be reduced.

IV. Backsliding & the Limits of a Purely Institutional Account

Some of the most salient democratic breakdowns of the Twentieth Century – such as those associated with Hitler and Mussolini – involved broad suspensions of civil liberties, the explicit prohibition of competing political parties, widespread jailing of opposition leadership, the use of substantial force to manipulate electoral outcomes, and – ultimately – the outright cancelation of elections. This strategy involves considerable risk for rulers, but Hitler and Mussolini were able to overcome the threat of popular resistance, in part, via their paramilitary forces, which were at least as large as the state army. The institutional account helps explain why, without access to the kind of military power that Hitler and Mussolini privately controlled, few leaders undermine democracy through the outright cancelation of elections.

even a rough commitment to similar types of rights, it may be difficult to prevent government overreach. I return to this point, and its relation to the Rawlsian approach, in Section VI.

Nevertheless, executive takeovers have accounted for four out of five democratic breakdowns since 2000.[46] Important, if more subtle, channels through which such takeovers occur include: making it more difficult for the opposition's supporters to vote, preventing the media from reporting critically on the government (e.g., by protecting the executive from "defamation"), and undermining judicial independence. These channels are attractive to incumbents because they are less likely to trigger widespread public outrage and resistance than straightforwardly canceling, or refusing to comply with, elections.

Still, they *do* involve the violation of rights. In light of the preceding section's analysis, it may seem surprising that executives pursue such strategies and that, when they do, they are not reliably thrown out of office. After all, these are precisely the kinds of strategies that enumerated rights and an independent judiciary are meant to foreclose. How, then, is such backsliding possible, consistent with the mechanisms elaborated in the previous section? And why has backsliding *recently* emerged as a threat in at least some relatively wealthy democracies—places previously thought to be relatively immune to such developments?

Promising answers to these questions can be inferred from the explanation of institutional stability described above—in particular, successfully undermining democratic institutions requires preventing citizens from mounting coordinated opposition to the government's efforts. Consider, then, two strategies that – along with *stealth* – allow governments to undermine democratic institutions without triggering such a response:

1. *Packaging*: in a sufficiently polarized environment, leaders may be able to package together undemocratic procedural reforms with substantive policy proposals that are strongly favored by their supporters, thus winning office by taking advantage of the (perceived) sheer unacceptability of the opposition's substantive policy positions.[47] If I am convinced that climate change is an existential threat and that the opposition party is unwilling to address it, then I may be willing to support a candidate who has a record of addressing climate change *even if*, much to my dismay, the candidate also has a record of preventing the media from reporting on the government. Here, the coarseness of democratic accountability, along with frictions that prevent a perfectly competitive political environment, can provide enterprising leaders with a valuable opportunity.

[46] Svolik 2019, p. 20–21.

[47] Svolik 2019; Graham and Svolik 2020; Şaşmaz, Yagci, and Ziblatt 2022.

2. *Uncertainty*: Citizens may prefer living under democratic institutions, but nevertheless be willing to support attempts to undermine such institutions if they are convinced that their political adversaries are prepared to do so.[48] In this kind of case, citizens face an assurance game. The structure of this situation presents an opportunity to a governing party that seeks to pursue executive aggrandizement. If it can convince its supporters that their adversaries intend to do away with democratic institutions, then those supporters may support executive aggrandizement as a preemptive measure—*even though* they (and their counterparts) strongly prefer living under democratic institutions. There is, indeed, empirical evidence showing that the expectation that the other side will find a way to undermine democratic institutions weakens the willingness of citizens to oppose their own party's attempts to do so.[49] From this perspective, it is not surprising that leaders who themselves wish to undermine democracy are so often at pains to emphasize (accurately or not) that their rivals are engaged in voter fraud and other kinds of undemocratic behavior.

Packaging and uncertainty describe central mechanisms through which government leaders manage to undermine democratic institutions—mechanisms that are available even *within* the kind of institutional environment described in Section III. They, in effect, describe ways in which the government can play factions off of one another and thereby undermine the likelihood of joint resistance.

Why, though, has backsliding emerged recently as a threat in countries that previously seemed immune? The most important part of an answer to this question is that the potential of these strategies, along with that of stealth, depends on broader circumstances. Critically, polarization – which is increasing in many of today's democracies – makes it easier for governments to unravel democratic institutions through each of the three identified strategies (stealth, packaging, and uncertainty).[50]

- Stealth is easier in a polarized environment because media coverage is and/or is viewed as partisan, which prevents the incumbent's anti-democratic actions from being broadly understood.[51]

---

[48] *Cf.* Cohen 2022; Pevnick 2023.

[49] Simonovits, McCoy and Littvay 2022; Braley et al. 2023.

[50] I understand a highly polarized setting to be one in which parties are far apart on policy issues and party supporters are tightly clustered around the partisan mean (Poole and Rosenthal 2011, p. 105).

[51] E.g., Pierson and Schickler 2024, ch. 8.

- In the absence of polarization, the packaging strategy would be difficult to execute, since the opposition would converge on a similar substantive platform without the deviations from democratic procedural norms.[52]
- It is easier to take advantage of uncertainty in a polarized environment because citizens in such an environment are more likely to be suspicious of the preferences and intentions of their political adversaries.[53]

Thus, polarization importantly enables executive aggrandizement by making it difficult for citizens to recognize that such efforts are underway, as well as by complicating efforts to coordinate resistance.[54] It is no accident, then, that backsliding has emerged as a threat to democratic societies as they have become more polarized.[55]

A subsidiary factor contributing to today's greater risk of backsliding seems to be more general knowledge, among political leaders, of the strategies available for successfully undermining democratic institutions in a piecemeal fashion. When populist leaders in one country find ways to take advantage of packaging strategies to undermine democratic institutions, that success is not lost on their counterparts. In this sense, leaders have become more sophisticated in their attempts to undermine democratic institutions. As one diplomat says, "Today, only amateurs steal elections on election-day."[56]

In sum, the discussion in this section revealed that – even given elections, enumerated rights and an independent judiciary – democracy cannot be expected to be reliably stable. Instead, as is increasingly understood by political leaders, there are strategies that allow for democracy to be undermined even in the face of these institutions—at least under polarized circumstances.

V. MITIGATING THE RISK OF BACKSLIDING

Having described the strategies that underlie successful backsliding, this section (a) explains how a constitutional consensus can, in the presence of proper institutions,

---

[52] It may be more difficult for parties to converge in a polarized setting if, for example, candidates need to raise money from, or ensure the turnout of, a more extreme base (Kujala 2019).

[53] E.g., Mason 2018.

[54] While this may indicate that a consensus on conceptions of justice can contribute to stability, the pathways through which it can do so are different from the one relied on by the Rawlsian account. I elaborate on this point in Section VI.

[55] *Cf.*, Levitsky and Ziblatt 2018; McCoy, Rahman and Somer 2018.

[56] Quoted in Bermeo 2016, p. 8.

help mitigate the threat of backsliding and (b) describes some institutions that can – together with elections, enumerated individual rights, and an independent judiciary – reduce the risk of backsliding.

A. The Emergence and Role of a Constitutional Consensus

In Section II, I argued that accounts of democratic stability rooted in the idea of a constitutional consensus are insufficient because they (1) lack an explanation of the mechanisms by which such a consensus could emerge and persist, and (2) fail to recognize that important mechanisms that drive backsliding can operate even in the presence of a constitutional consensus, which raises questions about the precise contribution that such a consensus can make to democratic stability. However, the institutional account allows us to address these lacunae and to identify the important role that a constitutional consensus *can* play.

What might cause a constitutional consensus to emerge and persist? It is plausible that, as many citizens experience the salutary effects of elections, enumerated rights, and an independent judiciary, they will come to value such institutions because they (1) provide a reasonably fair way of settling inevitable first-order disagreements, (2) provide some assurance that citizens' most basic interests will be protected, and (3) facilitate peaceful and mutually beneficial social cooperation. For these reasons, citizens may value such institutions *even when* they do not get the first-order policy outcomes that they most prefer.[57] Thus, the institutional account may be able to help explain how a constitutional consensus can emerge and persist.[58]

Is there an explanation of how a constitutional consensus can discourage backsliding that is consistent with recognizing its limited capacity to prevent stealth? When a constitutional consensus exists, particularly if citizens place significant weight on the importance of compliance with democratic institutions, leaders face a greater cost for eroding democratic institutions because the consensus forces them to worry that even those who otherwise support them on first-order policy grounds will join with the opposition if they attempt to erode democratic institutions. Thus, while it cannot *guarantee* that attempts at backsliding fail (citizens may still place greater weight on their first-order policy preferences), a constitutional consensus makes it *more difficult*

[57] One might object that the emergence of such a consensus itself requires background agreement on questions of justice—since, otherwise, citizens will disagree about whether the relevant institutions are, in fact, valuable. Yet, people can retain quite different ideals of justice while agreeing on (1), (2) and (3).

[58] *Cf.*, Barry 1970, p. 94; Weingast 1997, p. 253.

for leaders to pursue packaging and uncertainty-based strategies. The two accounts are, therefore, complementary: the incentive-based stability on which the institutional account turns *encourages, and can be importantly reinforced by*, a constitutional consensus.

B. Institutions for Mitigating Backsliding

As I have noted, understanding what facilitates democratic stability, as well as what allows leaders to interfere with it, should help us think about how to mitigate the threat of backsliding.

The three types of backsliding strategies that I described above – stealth, packaging, and uncertainty – may seem, on first glance, relatively disparate. However, the institutional account reveals their shared underlying similarity: they all make it more difficult for citizens to respond in a coordinated fashion to government overreach. In other words, these strategies allow governments to bypass the mechanisms that, under favorable conditions, facilitate democratic stability. While evaluation of specific reform proposals will be context dependent, this account encourages reformers to identify measures that will help citizens to (1) recognize attempts at government overreach and then (2) respond to such attempts in a coordinated fashion.

Begin with helping citizens to recognize attempts at government overreach. Since a fundamental role for the judiciary is to make stealth unavailable to the government by reliably publicizing its actions, stealth strategies will be most effective when paired with attempts to undermine citizens' faith in the judiciary. The government can do this in two related ways. First, the executive can *politicize* the judiciary, such that – although courts do provide an independent evaluation of the government's behavior – citizens do not view them as doing so. Second, the executive can *capture* the judiciary by filling it with reliable supporters, such that its rulings are not, in fact, reliable indicators of whether rights have been violated. The account therefore emphasizes the importance of preventing the politicization and capture of the judiciary. It may be that the traditional means of doing so (e.g., long terms for judges) need to be supplemented by more inventive arrangements (e.g., allowing the judiciary itself to determine which judges will sit on high courts). Also important from this vantage point are other institutions (such as independent election monitors and non-partisan media sources) that can reliably monitor and convincingly publicize the government's actions.

A second, less well-recognized, way to help facilitate citizens' awareness of attempts at government outreach is to remove legal power over the design of electoral

competition from the hands of contestants and their partisan allies.[59] Elected officials often have the power to design the terms of political competition. They can pass legislation that dictates: the rules of campaign finance, the rules for casting a countable ballot (e.g., should mail-in ballots be permitted? Is voter registration required?), the design of electoral districts, and so forth. When elected officials have a legal claim to design the rules of political competition, the question of whether their specific proposals count as eroding democratic institutions becomes a matter of partisan debate and legitimate disagreement. However, the design of political competition can be taken out of competitors hands, as when citizens vote directly on such matters or they are delegated to the judiciary or an independent body (such as a commission charged with designing electoral districts). Taking legal control over the relevant decisions out of the government's hands makes efforts to manipulate such processes more transparent and less subject to partisan disagreement. It also reduces the anxiety associated with *uncertainty* by offering citizens some reassurance that their political adversaries will not, even upon taking office, be able to undermine democratic institutions—thus reducing their own willingness to, pre-emptively, support such attempts.[60]

As important as it is to make attempts at government overreach publicly known, such efforts are insufficient—we have seen, after all, that efforts to erode democratic institutions can succeed even when citizens are aware that they are occurring (as in the *packaging* and *uncertainty* examples). In the face of these strategies, it is also important to find ways to increase the relative weight that citizens attach to democratic norms and institutions. This makes packaging strategies more difficult to execute and, if publicly known, undermines the prospects of the kind of preemptive sacrifice of democratic institutions associated with the uncertainty strategy. An obvious approach, which may increase the relative weight put on procedural considerations, is to limit polarization, so that there is less, in terms of substantive policy, at stake in elections. Unfortunately, given widespread empirical disagreement about the causes of polarization,[61] it is difficult to say, with conviction, what may be useful here, though familiar institutional proposals – such as ranked choice voting, open primaries, and proportional electoral systems – may help, in large part, by advantaging more moderate candidates.

While the preceding paragraphs provide some direction in thinking about how to protect democracies against backsliding, I do not mean to suggest that there is a set

[59] See McCoy and Somer 2019, 261 for a brief, related discussion.

[60] This paragraph draws on discussion in Landa and Pevnick 2025, ch. 6.

[61] E.g., McCarty 2019, ch. 5.

of institutional recommendations that can, with certainty, forestall backsliding. To the contrary, given what is at stake from the perspective of ambitious officeholders, it will be an ongoing challenge to effectively publicize, and then motivate coordinated opposition in response to, actions that undermine democratic institutions. The more modest hope is that, because the account is rooted in a compelling explanation of what makes democratic stability possible, it can help us to better understand democratic backsliding and, then, to clearly identify the kinds of steps that can mitigate the risk of it.

VI. BACK TO THE RAWLSIAN APPROACH

I have now argued (1) that the Rawlsian account is not well-positioned to help us to understand democratic stability or threats to it because the mechanism on which it relies is incompatible with due regard for the burdens of judgment and (2) that an institutional alternative can help us to better understand democratic stability and its vulnerabilities. This section considers two possible Rawlsian responses. The first suggests that the institutional account implicitly relies on an overlapping consensus and, so, does not actually deviate from the Rawlsian account. The second insists that even if the institutional account provides a superior framework for conceptualizing backsliding, it can – at best – underwrite a *modus vivendi*, which is problematic for the pragmatic and normative reasons introduced at the outset.

A. Does the Institutional Account Implicitly Rely on an Overlapping Consensus?

In light of the article's discussion, it is tempting to view the Rawlsian perspective as a kind of red herring—it focuses attention on a mechanism that cannot be expected to yield democratic stability under the relevant conditions and, in so doing, it diverts attention from more relevant considerations.

This way of glossing the point may, however, seem too strong, for even on the institutional account, one can identify (at least) two important ways in which a narrowing of disagreement about questions of justice can contribute to democratic stability. First, as we have seen, polarization is an important enabling condition for backsliding—it makes it easier to pursue each of the strategies described above. It could be, though, that having an overlapping consensus reduces polarization and, thereby, undercuts the capacity of executives to pursue such strategies.[62] Second, if

[62] It "could" do so because what is relevant from the perspective of backsliding is whether there is polarization with respect to preferred *policies*. Yet, one could – in principle – agree

there is a complete absence of agreement on what constitutes important rights, this will make it difficult to enumerate rights in a way that will enlist the shared response to government overreach that the institutional account requires.[63] Here, then, are two ways that a kind of overlapping consensus could, even on the institutional account, contribute to democratic stability.

However, identifying ways in which an overlapping consensus might contribute to democratic stability does not vindicate the Rawlsian account, for the underlying mechanisms at stake here are different. Shared conceptions of justice matter here, not because they help sustain a shared sense of reciprocity (as in the Rawlsian account), but rather because they indirectly facilitate a unified response to an executive bent on disabling democratic institutions. This reliance on different mechanisms is important for two reasons. First, as we saw in Section I, the mechanisms underlying the Rawlsian account are problematic—the reciprocity-based explanation of democratic stability does not seem like it could be operative given a realistic account of the level of disagreement about justice that is likely to emerge in a democratic society. Second, as we saw in Section V, with different underlying mechanisms, comes a different understanding of what leads to democratic backsliding and, so, a different way of thinking about the steps that can be taken to mitigate the risk of it. Thus, even if there is some role for shared views on the article's account, that does not imply that it is somehow parasitic on the Rawlsian account—instead, the underlying mechanisms enabling stability are dramatically different.

B. Is Institutional Stability Tantamount to a *Modus Vivendi*?

At this juncture, some readers are likely to think that even if the account that I have described provides a more satisfactory perspective for thinking about backsliding and democratic stability, it is nevertheless tantamount to a mere *modus vivendi*. This may seem concerning because, as we have seen, political stability rooted in *modus vivendi* has worrisome pragmatic and normative implications.

---

on a particular conception of justice, even while being polarized with respect to appropriate policies. There is vast disagreement among Rawlsians, for instance, about what policies, stretching from *laissez faire* to socialism, the two principles of justice require. So, agreement on a conception of justice does not seem to rule out the kind of staunch polarization on policy questions that enables the packaging and uncertainty strategies. It is possible, though, that it renders such polarization less likely.

[63] Note, however, that those who disagree substantially on broader questions of justice may nevertheless agree on a rough list of important liberties—think again of Rawls and Friedman.

Rawls worries, pragmatically, that a *modus vivendi* will collapse with a significant change in the underlying distribution of power. However, the point of the article's account is that – under the right conditions – there is no incentive for parties to abandon democratic institutions, even given considerable shifts in the underlying distribution of power. The discussion of elections' self-enforcing characteristics showed, for instance, that even officials who expect to be voted out of office often have very strong incentives to hold those elections and to comply with the results. Thus, partly by focusing on the longer run incentives facing key players, the article's position largely sidesteps the key pragmatic concern that Rawls associates with *modus vivendi* stability.[64]

One should be careful not to claim too much here. As the preceding discussion of backsliding makes clear, even well-designed democratic institutions cannot be regarded as *impervious* to breakdown. In addition to the vulnerabilities described above (i.e., stealth, packaging, and uncertainty), exogenous events could destabilize an institutional equilibrium. Yet, this risk of destabilization also exists under stability for the right reasons. Even a consensus among ordinary citizens on a particular conception of justice would not prevent the backsliding associated with "stealth" strategies. Less obviously, it would not necessarily prevent "packaging" strategies since people may agree on a conception of justice while disagreeing substantially about the policies that best satisfy that conception. (Note, again, the wide disagreement among Rawlsians about the policies that best satisfy Rawls's two principles.) Furthermore, some exogenous event could highlight the importance of a set of issues or concerns that had not previously been salient, and this could cause some people to doubt, and eventually abandon, their previous conception of justice, thereby upending the existing consensus.[65] From a pragmatic perspective, then, the article's account should not be regarded as fragile in the way that Rawls often portrays a *modus vivendi* and it may, perhaps, not even be importantly inferior – from that perspective – to stability for the right reasons.

Turning now to normative considerations, it is important to recognize that the account described above is compatible with more than compliance via the tacit threat of "brute force."[66] As described above, appropriate institutions may give rise to a constitutional consensus, which emerges because citizens recognize the value of society's central political institutions. Insofar as citizens embrace a society's procedures for making substantive decisions, they may view themselves as having

[64] *Cf.*, Basu 2021.

[65] Rawls 1999, p. 400.

[66] Rawls 1993, p. 171.

powerful (procedural) reasons to accept policy decisions, even when they view those decisions as substantively mistaken.[67]

Stability for the right reasons requires that "the reasons from which citizens act include those given by the account of justice they affirm."[68] Consider two ways to read this requirement. One possibility is that a system characterized by institutional stability and the resulting constitutional consensus *satisfies* the requirement because, even though some citizens disagree with the winning tax policy (for example), they nevertheless see it as emerging from a set of procedures that is endorsed by their conception of justice. On this reading, the account satisfies a necessary condition for stability for the right reasons and, so, may count as providing *a form of* such stability.[69]

A second possibility, though, is that stability is not normatively attractive unless all citizens live under *policies* whose value they accept. This reading would prevent us from seeing the article's account as facilitating a form of stability for the right reasons. Yet, even the Rawlsian account seems unable to satisfy this standard, since even if citizens manage to converge on a consensus on a particular conception of justice, there will still be disagreement about the *policies* that best satisfy that conception. The normative value that this interpretation associates with stability for the right reasons does not, then, appear consistent with a proper recognition of the burdens of judgment. The upshot is that it is surprisingly difficult to identify a tenable normative advantage associated with Rawlsian stability.

VII. CONCLUSION

Given recent events, democratic theorists are increasingly turning their attention to questions of democratic fragility and backsliding. While much of this work operates within a Rawlsian framework, we have seen that the reciprocity-based mechanism underlying the Rawlsian account of stability is incompatible with a proper appreciation of the burdens of judgment. Since it cannot explain – under realistic behavioral

[67] As I stressed above, these institutions can also provide reasons for compliance even to citizens who do not have an effective sense of justice and, so, either do not recognize, or do not give adequate weight to, these reasons. I take this to be a necessary feature of a successful account of stability, since – even under the most favorable plausible assumptions – there will be such individuals, and they will occasionally gain access to power.

[68] Rawls 1993, p. xlii.

[69] It "may" count (as opposed to "does" count) as providing a form of stability for the right reasons because – as a referee rightly points out – it could be that stability for the right reasons requires other conditions to be met, as well.

assumptions – how democratic societies could be stabilized, it also cannot explain how to mitigate threats to such stability.

I have, therefore, described an institutional alternative to the Rawlsian framework for thinking about problems of democratic stability. This alternative framework stresses the importance of institutions that give key actors self-interested reasons for compliance. This position's key mechanisms are operative under the conditions of disagreement that we should expect in democratic societies, and, in part for that reason, it provides a more realistic explanation of how democratic stability is possible than does the Rawlsian alternative. Understanding how democratic institutions can, under optimal circumstances, be stable also helps one to understand how the strategies that leaders use to erode democratic institutions work. This makes it easier to see what institutions would need to accomplish in order to mitigate the risk of backsliding—notably, making clear to citizens that such efforts are underway and lowering the costs to citizens of responding in a unified manner.

As democratic theorists increasingly turn to thinking about problems of democratic fragility, the institutional account promises, then, a more useful framework for understanding, and thinking about how to prevent, backsliding. It also, I hope, draws attention to virtues of stability possessed by traditional democratic institutions—virtues that are often overlooked in broader debates in democratic theory.

ACKNOWLEDGMENTS

For helpful discussion, I thank the participants at workshops at Stanford University, Bowling Green State University, and The Australian National University. I am particularly grateful to Brian Coyne, Bob Goodin, George Klosko, Michael Neblo, Josh Ober, Armando Jose Perez-Gea, Ana Tanasoca, and Kevin Vallier for written comments. The article also benefited enormously from a series of terrific conversations with Tony Taylor and Paul Weithman about the idea of stability in Rawls's work. I also thank Josh Cohen, who presented a characteristically thoughtful article on related themes at NYU in December of 2022. This article is, in part, an attempt to elaborate on the disagreement that the two of us had on that occasion. Finally, I have spent an enormous amount of time talking with Dimitri Landa about themes that are related to this article over the last several years, and these conversations have deeply shaped my thinking. Thus, while he is not responsible for the article's shortcomings, any merit that it may have surely reflects, in good part, his influence.

COMPETING INTERESTS

The author declares that he has no competing interests.

REFERENCES

Acemoglu, Daron and James A. Robinson. 2006. *Economic Origins of Dictatorship and Democracy*. New York: Cambridge University Press. https://doi.org/10.1017/CBO9780511510809

Baier, Kurt. 1989. Justice and the aims of political philosophy. *Ethics*, 99: 771–790. https://doi.org/10.1086/293121

Barry, Brian. 1970. *Sociologists, Economists, and Democracy*. Chicago: University of Chicago Press.

Barry, Brian. 1995. John Rawls and the search for stability. *Ethics*, 105: 874–915. https://doi.org/10.1086/293756

Basu, Jacqueline. 2021. Cooperative capacities of the rational: revising Rawls's account of prudential reasoning. *American Political Science Review*, 115: 967–981. https://doi.org/10.1017/S0003055421000101

Berman, Sheri. 2021. The causes of populism in the West. *Annual Review of Political Science*, 24: 71–88. https://doi.org/10.1146/annurev-polisci-041719-102503

Bermeo, Nancy. 2016. On democratic backsliding. *Journal of Democracy*, 27(1): 5–19. 10.1353/jod.2016.0012

Braley, Alia; Gabriel S. Lenz; Dhaval Adjodah; Hossein Rahnama; and Alex Pentland. 2023. Why voters who value democracy participate in democratic backsliding. *Nature Human Behaviour*, 7: 1282–1293. https://doi.org/10.1038/s41562-023-01594-w

Carugati, Federica. 2020. Democratic stability: a long view. *Annual Review of Political Science*, 23: 59–75. https://doi.org/10.1146/annurev-polisci-052918-012050

Claassen, Christopher. 2020. Does public support help democracy survive? *American Journal of Political Science*, 64: 118–134. https://doi.org/10.1111/ajps.12452

Cohen, Joshua. 2022. Reflections on democracy's fragility. Unpublished manuscript. Available at https://www.law.nyu.edu/node/38512

Easton, David. 1965. *A Systems Analysis of Political Life*. New York: Wiley.

Fearon, James D. 2011. Self-enforcing democracy. *Quarterly Journal of Economics*, 126: 1661–1708. https://doi.org/10.1093/qje/qjr038

Friedman, Milton. 1991. *Capitalism and Freedom*. Chicago: University of Chicago Press.

Graham, Matthew H. and Milan W. Svolik. 2020. Democracy in America? Partisanship, polarization, and the robustness of support for democracy in the United States. *American Political Science Review*, 114: 392–409. https://doi.org/10.1017/S0003055420000052

Habermas, Jürgen. 2001. *The Postnational Constellation: Political Essays*, ed. and trans. Max Pensky. Cambridge, MA: MIT Press.
Haggard, Stephan and Robert Kaufman. 2021. *Backsliding: Democratic Regress in the Contemporary World*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108957809
Klosko, George. 1993. Rawls's "political" philosophy and American democracy. *American Political Science Review*, 87: 348–359. https://doi.org/10.2307/2939045
Kogelmann, Brian. 2017. Justice, diversity, and the well-ordered society. *Philosophical Quarterly*, 67: 663–684. https://doi.org/10.1093/pq/pqw082
Krishnarajan, Suthan. 2023. Rationalizing democracy: the perceptual bias and (un) democratic behavior. *American Political Science Review*, 117: 474–496. https://doi.org/10.1017/S0003055422000806
Kujala, Jordan. 2019. Donors, primary elections, and polarization in the United States. *American Journal of Political Science*, 64: 587–602. https://doi.org/10.1111/ajps.12477
Landa, Dimitri and Ryan Pevnick. 2025. *Representative Democracy: A Justification*. Oxford: Oxford University Press.
Levitsky, Steven and Daniel Ziblatt. 2018. *How Democracies Die*. New York: Crown.
Levy, Jacob T. 2007. Federalism, liberalism, and the separation of loyalties. *American Political Science Review*, 101: 459–477. https://doi.org/10.1017/S0003055407070268
Lipset, Seymour Martin. 1959. Some social requisites of democracy: economic development and political legitimacy. *American Political Science Review*, 53: 69–105. https://doi.org/10.2307/1951731
Luo, Zhaotian and Adam Przeworski. 2023. Democracy and its vulnerabilities: dynamics of democratic backsliding. *Quarterly Journal of Political Science*, 18: 105–130. http://dx.doi.org/10.1561/100.00021112
Mason, Lilliana. 2018. *Uncivil Agreement: How Politics Became Our Identity*. Chicago: University of Chicago Press.
McCarty, Nolan. 2019. *Polarization: What Everyone Needs to Know®*. Oxford: Oxford University Press.
McCoy, Jennifer; Tahmina Rahman; and Murat Somer. 2018. Polarization and the global crisis of democracy: common patterns, dynamics, and pernicious consequences for democratic polities. *American Behavioral Scientist*, 62: 16–42. https://doi.org/10.1177/0002764218759576
McCoy, Jennifer and Murat Somer. 2019. Toward a theory of pernicious polarization and how it harms democracies: comparative evidence and possible remedies. *The Annals of the American Academy of Political and Social Science*, 681(1): 234–271. https://doi.org/10.1177/0002716218818782
Ober, Josiah. 2017. *Demopolis: Democracy Before Liberalism in Theory and Practice*. New York: Cambridge University Press.
Pevnick, Ryan. 2023. Immigration, backlash, and democracy. *American Political Science Review*, 118: 1–13. https://doi.org/10.1017/S000305542300028X

Pierson, Paul and Eric Schickler. 2024. *Partisan Nation: The Dangerous New Logic of American Politics in a Nationalized Era*. Chicago: University of Chicago Press.
Poole, Keith T. and Howard L. Rosenthal. 2011. *Ideology and Congress*. New Brunswick, NJ: Transaction.
Przeworski, Adam. 1991. *Democracy and the Market: Political and Economic Reforms in Eastern Europe and Latin America*. New York: Cambridge University Press.
Przeworski, Adam; Michael E. Alvarez; Jose Antonio Cheibub; and Fernando Limongi. 2000. *Democracy and Development: Political Institutions and Well-Being in the World, 1950–1990*. New York: Cambridge University Press.
Rawls, John. 1993. *Political Liberalism*. New York: Columbia University Press.
Rawls, John. 1999. *A Theory of Justice*. Cambridge, MA: Harvard University Press.
Rawls, John. 2001a. *Collected Papers*, ed. Samuel Freeman. Cambridge, MA: Harvard University Press.
Rawls, John. 2001b. *Justice as Fairness: A Restatement*. Cambridge, MA: Harvard University Press.
Şaşmaz, Aytuğ; Alper H. Yagci; and Daniel Ziblatt. 2022. How voters respond to presidential assaults on checks and balances: evidence from a survey experiment in Turkey. *Comparative Political Studies*, 55: 1947–1980. https://doi.org/10.1177/00104140211066216
Scheffler, Samuel. 2019. The Rawlsian diagnosis of Donald Trump. *Boston Review*, 12 February.
Simonovits, Gabor; Jennifer McCoy; and Levente Littvay. 2022. Democratic hypocrisy and out-group threat: explaining citizen support for democratic erosion. *Journal of Politics* 84: 1806–1811. https://doi.org/10.1086/719009
Svolik, Milan W. 2019. Polarization versus democracy. *Journal of Democracy*, 30(3): 20–32. https://dx.doi.org/10.1353/jod.2019.0039
Thrasher, John and Kevin Vallier. 2018. Political stability in the open society. *American Journal of Political Science*, 62: 398–409. https://doi.org/10.1111/ajps.12333
Waldner, David and Ellen E. Lust. 2018. Unwelcome change: coming to terms with democratic backsliding. *Annual Review of Political Science*, 21: 93–113. https://doi.org/10.1146/annurev-polisci-050517-114628
Weingast, Barry R. 1997. The political foundations of democracy and the rule of the law. *American Political Science Review*, 91: 245–263. https://doi.org/10.2307/2952354
Weithman, Paul. 2023. Rawls and Cohen, stability and justice. Unpublished manuscript.