The Distributional Effects of Minority-Protective Judicial Review

Justin Fox† Matthew C. Stephenson‡

June 8, 2011

Abstract

Constitutional theorists usually assume that minority-protective judicial review leads to outcomes more favorable to the protected minority and less favorable to the majority. Our analysis highlights three effects of judicial review that complicate, and sometimes undermine, this conventional wisdom. First, judicial review can induce a shift from a separating equilibrium—in which pro-majority leaders and pro-minority leaders pursue different policies—to a semiseparating or pooling equilibrium in which pro-minority leaders sometimes mimic pro-majority leaders by adopting the most anti-minority policy that the judiciary would uphold. Second, if judicial invalidation of anti-minority policies is probabilistic rather than certain, pro-majority leaders may propose even more extreme anti-minority policies in order to deter pro-minority leaders from mimicking. Third, if voters cannot directly observe policy outcomes, then minority-protective judicial review may create incentives for a leader to signal (or conceal) her type by provoking judicial reversal. These effects can sometimes nullify, or even reverse, the assumed relationship between minority-protective judicial review and pro-minority outcomes. When such reversal occurs, majoritarian democrats should favor minority-protective judicial review, while those concerned with protecting unpopular minorities should oppose it.

†Assistant Professor, Department of Political Science, Yale University, ISPS, P.O. Box 208209, New Haven, CT 06520. Email: justin.fox@yale.edu
‡Professor, Harvard Law School, Griswold 509, Cambridge, MA 02138. Email: mstephen@law.harvard.edu (corresponding author)
Constitutional judicial review, which empowers unelected judges to reject the decisions of elected legislatures and executives, raises what Bickel (1962) famously described as a “countermajoritarian difficulty.” After all, democratic political systems share a commitment to the principle of majority rule, partly on the grounds that policies (or leaders) supported by a majority are inherently more legitimate (Dahl 1989), and also on the more instrumental grounds that the regular elections ameliorate the principal-agent problem inherent in representative government, thereby producing more socially desirable policies (Fearon 1999; Riker 1982). For these and other reasons, many prominent constitutional theorists have urged the rejection or sharp curtailment of countermajoritarian judicial review (Tushnet 1999; Waldron 2006). At the same time, however, pure majority rule is normatively problematic. Even if one puts aside questions about whether voters understand their own interests, the ideal policy of the median voter may not maximize aggregate social welfare (Stiglitz 2000, pp. 237-241). Moreover, majoritarian policymaking may have undesirable distributional consequences—a “tyranny of the majority”—in which certain vulnerable minority groups consistently lose out (Guinier 1994). Majoritarian electoral pressures may also distort policymaking in directions that are undesirable, even for the majority (Canes-Wrone, Herron and Shotts 2001; Fox and Stephenson 2011). Many influential legal scholars see judicially enforced constitutional limits as a possible remedy for these pathologies; on this view, the countermajoritarian nature of appropriately designed judicial review is something to celebrate rather than condemn (Chemerinsky 2004; Dworkin 1985; Ely 1980). Important contributions to the political economy literature suggest similar conclusions (Maskin and Tirole 2004; Rogers and Vanberg 2007).

All sides of the normative debate over countermajoritarian judicial review make certain positive assumptions about the effect that such review is likely to have on policy outcomes. One such assumption is that courts are in fact often willing to block policies that adversely

---

1 Many theorists qualify this claim by identifying certain criteria that must be satisfied before majoritarian decisions have an intrinsic claim to legitimacy (Cohen 1997; Dworkin 1996)

2 Additional instrumental arguments for majority rule include the claim (derived from Condorcet’s jury theorem) that majority rule systems effectively aggregate dispersed information (Dahl 1989).
affect political minorities. Although some have challenged that assumption (Dahl 1957; Klarman 1998; Friedman 2009), we will provisionally accept the proposition that judicial review protects certain minority interests, at least some of the time. A second conventional assumption is that minority-protective judicial review leads to policy outcomes that are closer to those favored by the relevant minority group and further from those favored by the majority. This latter assumption—which undergirds both the standard critique of judicial review as antidemocratic and the standard defense of judicial review as a bulwark against majoritarian excesses—is natural and intuitive. As this paper will show, however, it may often be wrong. Although minority-protective judicial review sometimes benefits the minority at the expense of the majority, such review can sometimes hurt both the majority and the minority, or benefit both the majority and the minority, or affect the expected welfare of one group without affecting the other group’s expected welfare at all. In some cases, seemingly countermajoritarian judicial review may even benefit the majority at the expense of the minority.\(^3\) These results arise because judicial review—which, as we model it, imposes an outer limit on how anti-minority the final policy outcome may be—not only has the direct effect of prohibiting overly anti-minority policies, but may also have three indirect effects.

First, a judicially-imposed constraint on policy outcomes alters leaders’ ability and incentive to use their policy choices to conceal or reveal their true characteristics, and thereby to affect their popularity with voters. Leaders who share the majority group’s policy preferences have a reputational incentive to credibly signal this fact, while leaders who sympathize with

\(^3\)To be clear, we are not the first to argue that minority-protective judicial review may have these sorts of surprising effects. Most importantly, an established strain in the literature emphasizes that a full evaluation of the impact of judicial review on minority and majority welfare must consider not only the (implicit) distributional preferences of the judiciary (that is, the degree to which judicial review protects minority interests), but also the relative institutional competence of courts and other branches of government. Interestingly, some analyses in this vein assume that courts are more competent in assessing the connection between policies and outcomes, perhaps due to the fact that judicial review occurs later in time and in the context of a concrete case (Fleck and Hanssen 2010; Rogers 2001), while other analyses assume that courts are less competent, perhaps due to their lack of resources, expertise, and broad perspective (Fox and Stephenson 2011; Vermeule 2006). In a further development of this strain in the literature, Fleck and Hanssen (2010) have suggested that the distributional impact of countermajoritarian judicial review by better-informed courts may affect the majority’s willingness to enact policies that have positive expected net benefits for both the majority and the minority group. Our analysis differs from these prior contributions in that we do not assume that the judiciary and the elected leader have different levels of competence. This admittedly unrealistic simplification allows us to highlight effects of judicial review that derive from other sources.
the minority may have an incentive to conceal this information by mimicking the expected behavior of pro-majority leaders. Minority-protective judicial review can interfere with the pro-majority leader’s ability to signal her type, thereby leading pro-minority leaders to mimic more often. To illustrate, suppose that without judicial review, a pro-majority leader would adopt a policy so unfavorable to the minority group that a pro-minority leader would not be willing to mimic it. Under these circumstances, the pro-minority leader would instead adopt the minority group’s most-preferred policy. Now suppose that judicial review imposes a constraint on the pro-majority leader’s choice, forcing her to enact a policy somewhat more favorable to the minority faction. This reduces the cost to the pro-minority leader of mimicking the pro-majority leader; if this cost becomes small enough, the pro-minority leader will pool with the pro-majority leader, adopting a policy that is less favorable to the minority group than the pro-minority leader would have adopted otherwise.

Second, if there is uncertainty as to whether the judiciary will in fact invalidate policy proposals that impinge on minority interests, then judicial review may expand the set of anti-minority policies that the pro-minority leader would be willing to propose. Because judicial review means that the pro-minority leader may not have to live with the consequences of an extreme anti-minority policy, she is more willing to propose relatively extreme anti-minority policies if this would allow her to mimic the pro-majority leader. However, the pro-majority leader reacts to this by making an even more extreme anti-minority policy proposal, so that she can continue to distinguish herself from the pro-minority leader. The uncertainty of judicial review deters the pro-minority leader from mimicking such an extreme proposal. Although the court will sometimes invalidate the pro-majority leader’s extreme proposal, which makes the minority group better off than it would have been in the absence of judicial review, there will also be cases in which the court allows the pro-majority leader to implement this more extreme anti-minority policy, making the minority group worse off than it would have been in the absence of judicial review.

Third, if voters cannot directly observe policy, judicial approval or reversal may communicate information to voters about what policy the leader proposed, and (especially in the
case of judicial invalidation) what the final policy actually is. By picking a fight with the
courts—and thereby demonstrating to voters that the ultimate outcome is the most extreme
anti-minority policy the courts are willing to accept—a pro-majority leader can signal her


Thus minority-protective judicial review may sometimes lead pro-majority leaders to
take a more aggressively anti-minority position than they otherwise would. Paradoxically,
it is precisely when judicial review is not that stringent—when it prohibits only relatively


extreme policies—that this effect is most pronounced, and most likely to harm the interests
of both the majority and the minority groups. Furthermore, when judicial review is more


strongly countermajoritarian, even a pro-minority leader may have an incentive to pick a


fight with the court (pushing policy in as far an anti-minority direction as the courts will


allow) in order to conceal her type.


This paper uses a family of related formal models to illustrate these effects and to analyze
their implications. Part I develops a baseline model of policy choice in the absence of judicial
review. In this model, the incumbent leader may either share the policy preferences of the


majority group or else share the preferences of an unpopular minority group. In addition
to caring about policy, the leader also has an incentive to appear to share the majority’s
preferences. This baseline model allows us to isolate a particular set of political incentives
and to assess their effect on equilibrium political outcomes. Part II introduces judicial review,
modeled as a limit on the extent to which the policy outcome may harm the minority group’s


interests. Part II demonstrates the first of the three indirect effects sketched above: how
judicial review, by reducing the cost to the pro-minority leader of mimicking the pro-majority
leader’s policy choice, can induce a shift from a separating equilibrium to a semiseparating or
pooling equilibrium. Part III analyzes a variant of the model in which judicial invalidation of
anti-minority policies is probabilistic rather than certain. This analysis illustrates the second
of the indirect effects noted above: uncertain judicial review induces pro-majority leaders
to make even more extreme anti-minority proposals in order to distinguish themselves from
pro-minority leaders. Part IV then considers an alternative model in which voters cannot
observe the leader’s policy choice directly, but can observe whether the leader stayed within
the law or provoked judicial reversal. This model demonstrates the third indirect effects sketched above: how judicial review may create opportunities for a leader to signal (or conceal) her type by provoking conflict with the court. A brief conclusion follows, and an appendix provides proofs of all formal statements in the main text.

I. The Baseline Model

In this section, we develop a simple model of competitive politics with voter uncertainty regarding the politicians’ true preferences. Consider an elected leader responsible for enacting some policy, represented as a point in a one-dimensional policy space, \( x \in \mathbb{R} \). There are two groups in society, a majority faction \((M)\) and a minority faction \((m)\). The minority’s ideal policy is \( y_m = 0 \), while the majority’s ideal policy is \( y_M = 1 \). We might imagine that the minority faction is some traditionally disadvantaged ethnic or religious group. In this case, \( x = 0 \) might represent a policy of remedial affirmative action, while \( x = 1 \) might represent a principle of formal non-discrimination (“colorblindness”) that in practice tends to favor members of the majority group. Intermediate polices, \( x \in (0, 1) \), might represent more robust non-discrimination policies (including those that forbid facially neutral policies that have an adverse “disparate impact” on the minority group, but that stop short of affirmative action). Values of \( x > 1 \) are more extreme policies, such as those that overtly discriminate against the disadvantaged minority. Alternatively, we might think of the minority and majority factions in economic terms, for example as a wealthy elite and a working-class majority. In that case,

4Our baseline model is closely related to recent work by Acemoglu, Egorov and Sonin (2010). Their model (developed concurrently but independently from ours) seeks to explain seemingly excessive economic “populism” in countries where voters are suspicious of a wealthy elite. Although there are some important differences between our models, many of the principal assumptions and results are broadly consistent. One might therefore interpret our subsequent analysis of judicial review as complementary to their contribution, in that our paper considers the possible effects of introducing judicial review as a constraint on the sort of populism that they describe.
5Our model assumes that the leader is selected from a single constituency, that of the entire society. One could also apply our model to a smaller constituency, such as a legislative district or party primary. In such cases, the “majority” and “minority” factions would be defined as the relevant majority and minority in the relevant constituency, rather than the larger polity. However, a model that had multiple political actors, such as multiple legislators, would introduce substantial additional complexity into the model, particularly with respect to the attenuated connection between each elected representative’s action and the final outcome. We suppress that complexity here by focusing on a single leader elected from a polity-wide constituency.
$x = 0$ might represent a strict laissez-faire economic policy (most preferred by the wealthy minority), with higher values of $x$ representing greater levels of redistribution (either directly or through regulation). In this case, $x = 1$ would be the level of redistribution preferred by the working-class majority, while values of $x > 1$ denote expropriative policies that go beyond what even the workers who ostensibly benefit from redistribution would favor.

The leader has sincere policy preferences that may align either with the minority or with the majority. That is, the leader has a type, $\tau \in \{m, M\}$, where type-$m$ leaders prefer $x = 0$ and type-$M$ leaders prefer $x = 1$. The probability that $\tau = M$ is $p \in (0, 1)$; the probability that $\tau = m$ is $1 - p$. The voters do not know $\tau$, but $p$ is common knowledge. In other words, $p$ can be thought of as a representative voter’s prior belief about the leader’s type when the voter has taken into account all observable information about the leader before the leader selects policy $x$; such publicly observable information might include the leader’s partisan affiliation, prior statements and actions, and demographic characteristics. For now, we also assume that voters can observe the leader’s choice of $x$. The leader and each group incur a utility loss, denoted $L_\tau$, equal to the distance between the chosen policy and that player’s ideal policy, $L_\tau = |x - y_\tau|$.8

The leader also cares about her popularity. Let $\hat{p}$ denote the voters’ posterior estimate of the probability that the leader’s true sympathies lie with the majority faction (that is, that

---

6It might seem problematic to allow values of $p < \frac{1}{2}$, as such values imply that the probability of a pro-minority leader exceeds the probability of a pro-majority leader. That would indeed be problematic if the distribution of preferences among potential leaders were identical to the distribution of preferences in society, but this is not necessarily the case. If leaders are drawn from a more “elite” segment of the population, then it is plausible that leaders may sometimes be more likely to share the policy preferences of even a relatively small minority. Of course, regular elections are supposed to enable voters to select leaders who share their preferences (Fearon 1999), which might imply that $p$ should increase over time. But there is no a priori reason to suppose $p$ will converge to 1, particularly if turnover of individual leaders is rapid (say, due to term limits or retirement) and other, more durable forms of information (such as party affiliation) are only weakly correlated with a leader’s true sympathies as between the relevant minority and majority groups. Furthermore, politicians have a “bundle” of policy preferences, and voters typically must select from a limited number of candidates (Berry and Gersen 2008). The relevant policy choice, $x$, might be one on which even a leader who shares the majority’s preferences on most other dimensions might be thought to have pro-minority sympathies.

7It would therefore be a mistake to interpret type-$m$ leaders and type-$M$ leaders as members of different political parties. Rather, the partisan affiliation of the leader is incorporated into the prior, $p$.

8Note that this assumption implies risk neutrality.
\( \tau = M \). The higher \( \hat{p} \), the more popular the leader is. This follows from the assumption that every voter prefers a leader who shares her preferences, coupled with the fact that the majority group is by definition larger than the minority group. A voter may prefer a leader with similar preferences because such a leader is more likely to act in the voter’s interest on other issues (which may not be directly observable), or because the voter simply has an innate preference for a leader who shares her values. Likewise, the leader may care about her own popularity for either instrumental or intrinsic reasons. A popular leader may be more likely to win reelection or succeed in advancing some other policy initiative; leaders may also derive inherent satisfaction (“ego rents”) from popularity. We do not provide a fully micro-founded model of these aspects of the voters’ and leaders’ utility functions, but for the reasons sketched above we view these assumptions as reasonable. Because the leader cares about her popularity, she receives a reputational utility payoff of \( \hat{p}w \), where \( w \geq 0 \) is the weight the leader assigns to popularity.\(^9\) The leader’s net utility payoff, taking into account both policy and reputational concerns, is therefore \( \hat{p}w - L_\tau \).

Our use of the term “popularity” as a shorthand to describe the leader’s interest in \( \hat{p} \) is, however, potentially misleading. Each voter’s attitude toward the leader may depend not only on what the voter thinks about the leader’s true type (which, again, might correlate with the leader’s unobservable policy decisions), but also on what the voter thinks of the leader’s decision on the observable policy decision, \( x \). Thus one might think of the leader’s utility function as having three components rather than two: (1) the leader cares about policy (which leads the type-\( m \) leader to favor \( x = 0 \) and the type-\( M \) leader to favor \( x = 1 \)); (2) the leader cares about the popularity of her observed choice (which would lead both leaders to favor \( x = 1 \), the most popular policy); and (3) the leader cares about the voter’s assessment her true type, which also affects the leader’s popularity (and which

\( ^9 \)This posterior is a function of the primitives, the players’ equilibrium strategies, and the observed policy choice, but for convenience we suppress this additional notation and write the posterior simply as \( \hat{p} \).

\( ^{10} \)The value of \( w \) may reflect the relative sizes of the majority and minority factions, as well as the amount the leader cares about popularity relative to the particular policy decision at issue. For simplicity, we assume that \( w \) is the same for both leader types. It would be possible to extend the model so that each leader type \( \tau \) placed a different weight, \( w_\tau \), on popularity, but we do not pursue that extension here.
would lead each leader to favor whatever choice of \( x \) maximizes \( \hat{p} \), given the other players’ equilibrium strategies and beliefs). To keep things simple, our models formally incorporate only considerations (1) and (3) into the leader’s utility function. One can, however, interpret the leader’s policy preference as implicitly incorporating consideration (2) as well: for the pro-majority leader, both considerations (1) and (2) favor outcome \( x = 1 \), while for the pro-minority leader, we can reinterpret her ideal point, 0, as incorporating her optimal trade-off between the minority faction’s “true” ideal point (which would be redefined as \( y_m < 0 \)) and her interest in enacting the most popular policy, \( x = 1 \). On this interpretation, the parameter \( w \) should be interpreted not as the leader’s general interest in popularity relative to policy, but rather as her interest in a particular element of her popularity—the voter’s estimate of her true type—relative to the combined interest in both policy and other elements of her popularity. For expositional simplicity, though, we will continue to use the term “popularity” (or “reputation”) to describe the former interest, and the term “policy” to describe the latter two interests.

We solve for Perfect Bayesian Equilibria (PBE), with two refinements. First, we restrict attention to PBE in which the type-\( M \) leader plays a pure strategy, denoted \( x_M \).\(^{11}\) Second, we restrict off-path beliefs such that \( \hat{p} = 1 \) if \( \tilde{x} > x_M \) and \( \hat{p} = 0 \) if \( \tilde{x} < x_M \), where \( \tilde{x} \) denotes any \( x \) off the equilibrium path.\(^ {12}\) We will refer to a PBE consistent with these refinements simply as an “equilibrium.”

The policy incentives of the leaders are straightforward: all else equal, each leader prefers to enact her ideal point (\( x_\tau = y_\tau \)). However, leaders’ reputational interests may pull in a different direction. All else equal, the type-\( m \) leader would prefer to conceal her type, because type-\( m \) leaders are less popular when voters can confidently identify them. Formally, we

\(^{11}\)Such equilibria always exist, as the subsequent analysis demonstrates. For some parameterizations of the model, there are also equilibria in which both types of leader use mixed strategies. We do not consider such equilibria in the main analysis, though we note in the appendix how and why they may arise.

\(^{12}\)That is, if voters were to observe a policy even more anti-minority than the policy the pro-majority leader is expected to choose in equilibrium, the voters would assume that the leader is pro-majority rather than pro-minority, while if the voters were to observe a policy choice closer to the minority group’s interest than the policy the pro-majority leader is expected to choose in equilibrium, the voters would assume that the leader is pro-minority. Though there are some technical differences, this restriction limits the set of equilibria in a manner similar to the “intuitive criterion” (D1) developed by Cho and Kreps (1987).
can derive the pro-minority leader’s optimal strategy (treating $x_M$ as fixed and assuming equilibrium voter beliefs) as follows:

**Lemma 1**

(a) If $x_M \geq w$, then the type-$m$ leader selects her ideal point, $x_m = 0$.

(b) If $x_M \leq pw$, then the type-$m$ leader mimics the type-$M$ leader by selecting $x_m = x_M$.

(c) If $x_M \in (pw, w)$, then the type-$m$ leader plays a mixed strategy in which she selects $x_m = x_M$ with probability $\left( \frac{p}{1-p} \right) \left( \frac{w-x_M}{x_M} \right)$, and selects $x_m = 0$ otherwise.

To see the intuition for part (a), observe that if the type-$m$ leader chooses her ideal policy $x = 0$, and the voters (correctly) expect the type-$M$ leader always to choose some $x_M > 0$, then the type-$m$ leader’s net payoff is 0. (She gets her maximum policy payoff at $x = 0$, but her choice distinguishes her as type-$m$, resulting in a posterior $\hat{p} = 0$.) If the voters expected the type-$m$ leader to select $x_m = 0$, but she deviated by instead selecting $x_m = x_M$, then her net payoff would be $w - x_M$ (as she realizes a posterior popularity of $\hat{p} = 1$ rather than $\hat{p} = 0$, but incurs a policy loss of $x_M$). This net payoff is (weakly) negative if but only if $x_M \geq w$.

The intuition for part (b) is similar. If the type-$m$ leader pools by selecting $x_m = x_M$, then her expected payoff is $pw - x_M$. This is (weakly) greater than her net payoff from selecting her ideal point if but only if $x_M \leq pw$. When that condition holds, the type-$m$ leader would always mimic the type-$M$ leader. Perhaps most interesting and subtle is the case described in part (c), where $x_M \in (pw, w)$. In this case, there is no equilibrium pure strategy for the type-$m$ leader: if voters expected her always to play her ideal point, she would have a strict incentive to deviate by mimicking the type-$M$ leader; if voters expected her always to mimic, she would have a strict incentive to deviate by selecting her ideal point. Thus the only equilibrium is one in which the voters’ beliefs make the type-$m$ leader indifferent between these two options, and her mixed strategy induces such beliefs in the voters.

Now consider the incentives of the type-$M$ leader. The type-$M$ leader could achieve her maximum policy payoff by always selecting $x_M = 1$. If $w \leq 1$, then we know from Lemma 1 that this choice comes at no reputational cost: the pro-minority leader would
never mimic, so the pro-majority leader can perfectly distinguish herself by choosing her ideal point. However, if \( w > 1 \) then the pro-majority leader incurs a reputational cost if she selects her ideal point, because we know (from Lemma 1) that the pro-minority leader will mimic her at least some of the time. Thus, it is not an equilibrium for the pro-majority leader to select an \( x_M \in [1, w) \): for any such \( x_M \), the pro-majority leader could always increase her net payoff by adopting a slightly higher \( x \), which decreases her policy payoff by a trivial amount but (discontinuously) increases her reputational payoff. The incentive to deviate disappears only when \( x_M = w \), at which point the pro-majority leader perfectly distinguishes herself.\(^{13}\) These incentives can be formalized as follows:

**Lemma 2** In equilibrium, the type-\( M \) leader chooses \( x_M = \max\{w, 1\} \).

Using Lemmas 1 and 2, we can fully characterize the equilibria of the model as follows:

**Proposition 1** (a) If \( w \in [0, 1] \), there is a sincere separating equilibrium in which the type-\( M \) leader selects the popular policy \((x_M = 1)\), the type-\( m \) leader selects the minority-preferred policy \((x_m = 0)\), and voters’ posterior beliefs are \( \hat{p} = 0 \) if \( x < 1 \) and \( \hat{p} = 1 \) if \( x \geq 1 \). The expected policy losses for each group are \( E(L_m) = E(x) = p \) and \( E(L_M) = 1 - p \).

(b) If \( w > 1 \), there is an extremist separating equilibrium in which the type-\( m \) leader chooses the minority-preferred policy, \( x_m = 0 \), but the type-\( M \) leader chooses a policy, \( x_M = w > 1 \), that is more anti-minority than the majority group’s ideal policy. Voters’ posterior beliefs are \( \hat{p} = 0 \) if \( x < w \) and \( \hat{p} = 1 \) if \( x \geq w \). The expected policy losses for each group are \( E(L_m) = E(x) = pw \) and \( E(L_M) = 1 - p(2 - w) \).

Observe that in this model pro-majority leaders always perfectly distinguish themselves from pro-minority leaders, and pro-minority leaders always choose their own ideal points. In the situation described in Proposition 1(a), the pro-majority leader also selects her own

\(^{13}\)The type-\( M \) leader clearly has no incentive ever to select any \( x_M < 1 \) or \( x_M > \max\{w, 1\} \), as in both cases the type-\( M \) leader would incur a policy loss for no reputational benefit.
ideal point. Perhaps more interesting is the case described in part (b). In this case, the pro-majority leader adopts a policy that is too extreme even from the majority group’s perspective, such that both the majority and the minority would benefit if the pro-majority leader adopted a more moderate policy. Indeed, it is possible (if $w > 2$) that the policy selected by the pro-majority leader will be even worse for the majority group than the policy selected by the pro-minority leader. Yet the voters still “reward” the pro-majority leader for adopting this sort of extreme policy by giving her a popularity boost.\textsuperscript{14}

While our main focus is on the effect of judicial review, the result derived in Proposition 1(b) may have independent significance. For some issues, political leaders care much less about the issue itself than about how their stance on this issue affects their popularity (which would correspond, in our model, to situations in which $w$ is large). Also, in some circumstances voters in the majority are worried that leaders might be overly sympathetic to some unpopular minority (in terms of our model, $p < 1$). When these conditions hold, leaders who share the majority’s preferences will go out of their way to demonstrate their antipathy to the minority group, even to the point of enacting measures that go far beyond what the median voter (or the pro-majority leaders themselves) would actually prefer. That is, the preceding analysis suggests a possible rationalist explanation for why seemingly excessive anti-minority extremism is sometimes a successful political strategy. There are numerous examples that suggest the empirical plausibility of this dynamic. For instance, Acemoglu, Egorov and Sonin (2010) use a model similar to ours to explain the prevalence of “populist”

\textsuperscript{14}A seemingly unrealistic prediction of our model is that for very large values of $w$, the type-$M$ leader would adopt a policy that hurts all voters, while the type-$m$ leader would continue to implement her ideal point. Our model does indeed predict that this is what would occur for high values of $w$. There are two reasons, however, why the apparent empirical implausibility of this prediction does not necessarily undercut the validity of the insights produced by the model. First, as noted above, the parameter $w$ should be thought of not as the weight that the leader assigns to popularity generally, but rather as how much she cares about convincing the voters that she is really type-$M$. Extremely high values of $w$ might therefore simply be empirically implausible, unless the voters view the observable policy issue as (mostly) trivial, except as a tool for drawing inferences about the leader’s true type. Second, in our simple model, the majority faction is only worried about one minority faction, located to the majority’s “left” (in arbitrary spatial terms). In a richer, more realistic model, the majority faction might also worry that the leader might too sympathetic to another minority faction to the “right.” When both concerns are present, there will be some constraint on how far a type-$M$ could deviate in any direction. We defer formal analysis of that alternative model to future research; we note it here as a reason why extreme values of $w$ might not necessarily produce arbitrarily large deviations from the majority faction’s ideal point.
politics, particularly in contemporary Latin America, whereby politicians implement policies well to the left of the median voter (and bad for the economy as a whole) in order to convince the electorate that they are not secretly sympathetic to the wealthy elite. Schama (1989) suggests a similar dynamic may have been operative during the French Revolution, with revolutionary leaders embracing increasingly violent policies in order to signal that they were not secretly sympathetic to the royalists. This dynamic is also familiar in American politics. For example, historians have suggested that southern politicians in the mid-twentieth century staked out policy positions that were even more hostile to African-Americans than the median (white) voter would have preferred, in order to signal to these voters that the politicians were not “soft” on maintaining the racial hierarchy. Perhaps the most well-known (and notorious) illustration of this is George Wallace’s declaration, after losing the Alabama gubernatorial election in 1958 to a Klan-backed candidate, that “no other son-of-a-bitch will ever out-nigger me again” (Carter 1995, p. 96). The dynamic also seems present in contemporary “culture war”-style politics, in which politicians and party activists sometimes stake out positions on high-profile issues that seem too extreme not only relative to the median voter in the electorate, but even with respect to the median voter within their own party (Fiorina, Abrams and Pope 2011).

Even without this extremism effect (that is, even when both leader types implement their respective ideal points, as in Proposition 1(a)), one might be concerned about the welfare of the minority group; when anti-minority extremism of the sort predicted by Proposition 1(b) is present, one might worry about the adverse effects on the welfare of both groups. Those who share those concerns might be attracted to institutional solutions, such as legal constraints coupled with effective judicial review, that prevent leaders from enacting overly anti-minority policies. On the other hand, majoritarians might worry that such institutions would undermine basic democratic commitments. Indeed, from a majoritarian perspective, the more important problem highlighted by Proposition 1 is that pro-minority leaders often implement a policy that favors the minority group, rather than catering to the median voter. Those sympathetic to that perspective might be skeptical of judicial constraints
designed to protect the minority group, though perhaps more mild constraints that mitigate
the political tendency toward “excessive” anti-minority extremism (of the sort highlighted
in Proposition 1(b)) might be welcome. The remainder of this paper explores whether, or
under what conditions, these instincts about the effect of minority-protective judicial review
are accurate.

II. The Effect of Judicial Review

To investigate the impact of minority-protective judicial review, we modify the baseline
model by introducing a judicial review stage that follows the leader’s selection of the policy,
x. The court protects the minority group by imposing a limit, j, on how far the policy choice
may deviate from 0 (the position that most benefits the minority group).\textsuperscript{15} If the leader
selects an \( x \in [0, j] \), the court will uphold the leader’s policy choice; if the leader selects
an \( x > j \), the court will strike down the leader’s choice and implement \( x = j \) instead.\textsuperscript{16}
The limit \( j \) is common knowledge.\textsuperscript{17} We do not explicitly model the utility function of the
judiciary; our formulation is consistent with a range of assumptions about judicial behavior,
including both the assumption that the court sincerely enforces a constitutional provision
that prohibits any \( x > j \), and the alternative assumption that the court is an ideologically
motivated strategic actor with preferences that induce acceptance of any \( x \in [0, j] \) and
rejection of any \( x > j \). For simplicity, we will not treat the court as a separate strategic
player, but simply as an institutional feature that induces an upper bound, \( j \), on the policy
space. We also assume, in this section, that no leader would ever propose a policy \( x > j \),

\textsuperscript{15}No leader ever has an incentive to select any \( x < 0 \), so we do not model judicial review of such policies.
\textsuperscript{16}One can think of the judicial decision as invalidating only those aspects of the policy that induce a
distributional outcome above \( j \). The effect of judicial review, in other words, is to “trim” the unacceptable
parts of the government policy, while leaving the rest intact. Thus judicial invalidation moves the outcome
from \( x \) to \( j \), rather than to some alternative “status quo” default.
\textsuperscript{17}Note that our model is similar to Groseclose and McCarty (2001), who also study a setting involving a
proposer, a potential vetoer, and an audience of voters. In their model, however, voters know the preferences
of the proposer (Congress in their model) but are uncertain about those of the veto player (the President),
whereas in our model voters know the preferences of the veto player (the court) but are unsure about the
preferences of the proposer (the leader).
as this would have no effect on the outcome. Also, for expositional convenience, we will refer to judicial review as “counter-majoritarian” when \( j \in [0, 1) \), and we will refer to judicial review as “anti-extremist” when \( j \geq 1 \). In the former case, judicial review prevents the leader from adopting certain policies that would benefit the majority at the expense of the minority. In the latter case, judicial review precludes only extreme policies that would be disfavored (on policy grounds) by both the minority and the majority groups.

The following two propositions establish how the introduction of judicial review alters the equilibrium outcome of the model. Each of the propositions corresponds to one of the subcases (parts (a) and (b)) of Proposition 1.

**Proposition 2** Suppose \( w \in [0, 1] \), such that in the absence of judicial review, there would be a sincere separating equilibrium. (See part (a) of Proposition 1.) The introduction of (counter-majoritarian) judicial review has the following effects, depending on the value of \( j \):

(a) If \( j \in [w, 1) \), then judicial review shifts the equilibrium to a judicially-constrained separating equilibrium in which the type-M leader selects \( x_M = j \), the type-m leader selects \( x_m = 0 \), and voters’ posterior beliefs are \( \hat{p} = 0 \) if \( x < j \) and \( \hat{p} = 1 \) if \( x = j \). The expected policy losses for each group are \( E(L_m) = E(x) = pj \) and \( E(L_M) = 1 - pj \).

(b) If \( j \in (pw, w) \), then judicial review shifts the equilibrium to a judicially-constrained semiseparating equilibrium, in which the type-M leader selects \( x_M = j \) and the type-m leader plays a mixed strategy, selecting \( x_m = j \) with probability \( \frac{p_j}{1 - p_j} \) and selecting \( x_m = 0 \) otherwise. Voters’ posterior beliefs are \( \hat{p} = 0 \) if \( x < j \) and \( \hat{p} = \frac{j}{w} \) if \( x = j \). The expected policy losses for each group are \( E(L_m) = E(x) = pw \) and \( E(L_M) = 1 - pw \).

---

\(^{18}\)This assumption, which greatly simplifies the exposition in Part II, implies a slight modification of our earlier refinement of PBE, which posited that, for any off-path \( \hat{x} \), \( \hat{p} = 1 \) if \( \hat{x} > x_M \). Here, because any proposal \( x \geq j \) results in final outcome \( j \), and is therefore welfare- and information-equivalent, we impose the refinement that a voter’s posterior belief upon observing any \( x > j \) is equal to the voter’s posterior beliefs upon observing \( x = j \); that is, we assume \( \hat{p} \) is constant for all \( x \geq j \). However, this assumption would not be appropriate if there were a positive probability that the court might uphold a policy \( x > j \). We explore that alternative case in Part III. Furthermore, when the voters cannot observe policy directly, but can observe judicial invalidation, a leader might have an incentive to propose a policy that she knows for certain the court will invalidate, because doing so may convey additional information to the voters. We take up this possibility in Part IV.

\(^{19}\)We ignore cases in which \( j \geq \max\{w, 1\} \), as judicial review has no effect in such cases.
(c) If \( j \in [0,pw] \), then judicial review shifts the equilibrium to a judicially-constrained pooling equilibrium, in which both leader types select \( x = j \) and voters’ posterior beliefs are \( \hat{p} = 0 \) if \( x < j \) and \( \hat{p} = p \) if \( x = j \). The expected policy losses for each group are \( E(L_m) = E(x) = j \) and \( E(L_M) = 1 - j \).

Proposition 2(a) corresponds to the conventional wisdom: countermajoritarian judicial review does not affect the behavior of pro-minority leaders, but pro-majority leaders are at least somewhat constrained (forced to implement \( j < 1 \) rather than 1), and this constraint unambiguously benefits the minority group at the expense of the majority. Parts (b) and (c) of Proposition 2 complicate the standard intuition, however. To understand these cases, note that as the countermajoritarian court forces the type-\( M \) leader to accept a policy outcome of \( j \) rather than 1, the costs to the type-\( m \) leader of pooling decrease. Once these costs are sufficiently small, then (as Lemma 1 established) it is not an equilibrium for the type-\( m \) leader always to select her ideal point. Rather, she will (sometimes) mimic the type-\( M \) leader. Thus judicial review not only has a direct effect on policy choice (forcing the type-\( M \) leader to select \( j \) rather than 1), but it also has an indirect effect on policy choice, inducing the type-\( m \) leader to (sometimes) pool rather than separate. Therefore, although countermajoritarian judicial review leads to outcomes that are better for the minority (and worse for the majority) when a pro-majority leader is in office, such judicial review leads to outcomes that are worse for the minority (and better for the majority) when a pro-minority leader is in office, compared to the no review baseline.

How does this indirect effect alter the expected welfare of the minority and majority groups? First, if judicial review is sufficiently constraining (in particular, if \( j \leq pw \)), then the pro-minority leader always mimics the pro-majority leader, but in expectation the direct constraining effect of judicial review dominates: as \( j \) decreases from \( pw \) to 0, the expected policy outcome moves toward the minority’s ideal point and away from the majority’s. Thus the intuitive relationship between countermajoritarian judicial review and minority/majority welfare continues to hold. This is the case described in Proposition 2(c).
Perhaps more striking is the case described in Proposition 2(b). Here, the direct and indirect effects of countermajoritarian judicial review offset, such that the expected policy outcome is always the same \((pw)\). Thus, Proposition 2(b) reveals that strengthening the countermajoritarian constraint (that is, decreasing \(j\)) may not have any effect on the expected policy outcome or on the expected policy payoff to either group. The range of values over which changes in the stringency of judicial review have no effect (the \([pw, w]\) interval) is larger when voters expect most leaders to favor the minority (\(p\) small), and when leaders care more about popularity relative to policy (\(w\) large, but less than 1).\(^{20}\)

The next proposition considers the effect of judicial review in a case where the pro-majority leader, in the absence of judicial review, would take an extreme anti-minority position in order to signal her type. As in the preceding case, assessing the impact of judicial review may require consideration of both direct and indirect effects.

**Proposition 3** Suppose \(w > 1\), such that in the absence of judicial review, there would be an extremist separating equilibrium. (See part (b) of Proposition 1.) The introduction of judicial review has the following effects, depending on the value of \(j\):

(a) If \(j \in (1, w)\), then (anti-extremist) judicial review shifts the equilibrium to a judicially-constrained semiseparating equilibrium, in which the type-\(M\) leader selects \(x_M = j\) and the type-\(m\) leader plays a mixed strategy, selecting \(x_m = j\) with probability \(\left(\frac{p}{1-p}\right)\left(\frac{w-j}{j}\right)\) and selecting \(x_m = 0\) otherwise. Voters’ posterior beliefs are \(\hat{p} = 0\) if \(x < j\) and \(\hat{p} = \frac{j}{w}\) if \(x = j\). The expected policy losses for each group are \(E(L_m) = E(x) = pw\) and \(E(L_M) = 1 - pw(\frac{2}{j} - 1)\).

(b) If \(pw < 1\) and \(j \in (pw, 1]\), then (countermajoritarian) judicial review shifts the equilibrium to a judicially-constrained semiseparating equilibrium, in which the type-\(M\) leader selects \(x_M = j\) and the type-\(m\) leader plays a mixed strategy, selecting \(x_m = j\) with probability \(\left(\frac{p}{1-p}\right)\left(\frac{w-j}{j}\right)\) and selecting \(x_m = 0\) otherwise. Voters’ posterior beliefs

\(^{20}\)If \(w > 1\), then Proposition 3 rather than Proposition 2 would apply, and further increases in \(w\) would shrink rather than enlarge the range in which changes in \(j\) have no effect on expected welfare.
are $\hat{p} = 0$ if $x < j$ and $\hat{p} = \frac{j}{w}$ if $x = j$. The expected policy losses for each group are $E(L_m) = E(x) = pw$ and $E(L_M) = 1 - pw$.

(c) If $j \in [0, \min\{pw, 1\}]$, then (countermajoritarian) judicial review shifts the equilibrium to a judicially-constrained pooling equilibrium, in which both leader types select $x = j$ and voters’ posterior beliefs are $\hat{p} = 0$ if $x < j$ and $\hat{p} = p$ if $x = j$. The expected policy losses for each group are $E(L_m) = E(x) = j$ and $E(L_M) = 1 - j$.

The results in Proposition 3 largely track the results in Proposition 2; indeed, parts (b) and (c) of Proposition 3 are essentially identical to parts (b) and (c) of Proposition 2. Proposition 3(a), however, highlights an additional result. Recall that in the circumstances described in Proposition 3, without judicial review the pro-minority leader would choose her ideal point but the pro-majority leader would choose an extreme policy $x_M = w > 1$, which is suboptimal for both the majority group and the minority group. Proposition 3(a) implies not only that anti-extremist judicial review benefits the majority group, but (perhaps more surprisingly) that anti-extremist review has no effect on the expected policy payoff of the minority group. To see why, observe that anti-extremist judicial review has two effects: a direct constraining effect on the pro-majority leader’s choice (shifting it from $w$ to $j$) and an indirect effect on the pro-minority leader’s choice (inducing her sometimes to select $j$ rather than 0). Both of these effects shift policy closer to the majority group’s ideal point, but for the minority group, these effects offset in expectation. Thus our model indicates not only that anti-extremist judicial review may help the majority (an intuitive result that accords with much of the extant literature), but that such review, though nominally minority-protective, may in fact benefit only the majority group. Furthermore, Proposition 3 implies that although the majority group would prefer anti-extremist judicial review to countermajoritarian judicial review, under some circumstances the majority prefers countermajoritarian judicial review of the sort described in Proposition 3(b) to no review at all.²¹

²¹To illustrate, suppose $p = 0.5$, $w = 2$, and $j = 0.8$. Without judicial review, the majority group’s expected policy loss would be 1. With review, the majority group’s expected policy loss is 0.2.
The main insight of the analysis to this point is that minority-protective judicial review may have two types of effect on policy outcomes. The first and more obvious effect is that judicial review constrains the degree to which the final policy can deviate from the minority group’s ideal outcome. The second and more novel finding is that if \( j < w \), judicial review can induce a shift from a fully separating equilibrium (in which type-\( m \) leaders select \( x = 0 \) and type-\( M \) leaders select \( \max\{1, w\} \)) to a semiseparating or pooling equilibrium in which even pro-minority leaders sometimes adopt policies that disfavor the minority group. This latter effect means that the impact of ostensibly minority-protective judicial review on both minority and majority interests is more complicated than the conventional wisdom suggests. Although minority-protective judicial review weakly benefits the minority group on the policy dimension, for a potentially broad range of parameter values introducing or strengthening minority-protective judicial review may not alter the minority group’s expected policy payoff at all. To make essentially the same point from the perspective of the majority, although countermajoritarian judicial review may injure the majority’s interests by (sometimes) constraining pro-majority leaders, such review may also benefit the majority by inducing leaders who favor the minority to adopt more majoritarian policies.\(^{22}\) Moreover, anti-extremist judicial review will redound to the benefit of the majority rather than the minority, and even truly countermajoritarian judicial review may sometimes be better for the majority group (as well as the minority group) than no review at all.

The discussion of welfare effects so far has focused on the expected policy losses, \( L_M \) and \( L_m \). However, focusing narrowly on the policy outcome may be misleading, in that judicial

\(^{22}\)This result bears a superficial resemblance to the argument, first advanced by Thayer (1893), that judicial review reduces the degree to which elected politicians pay attention to constitutionally-protected values, including minority rights (Fox and Stephenson 2011; Rogers 2009; Tushnet 1999; Vermeule 2006). In our model, as in the Thayerian account, the fact that the judicial constraint prevents the enactment of policies that are too anti-minority may lead some leaders (in particular, pro-minority leaders) to adopt policies that are less minority friendly than they would have otherwise. The reason, however, is different. In the Thayerian account, judicial review acts as a safety net, making politicians more reckless (that is, less attentive to the impact of their policy choices on constitutionally-protected interests). In the model developed in Part II, the pro-minority politician shifts her policy choice not because she expects the court to bail her out, but because the judicial constraint on the pro-majority leader makes pooling attractive for the pro-minority leader. As we will see in Part III, when judicial invalidation of anti-minority policies is uncertain, our model produces something similar to the sort of bailout effect discussed by Thayer and others, though even in that case the dynamic is somewhat different from the conventional account.
review can affect the welfare of the majority and minority groups through at least two other channels. First, by (sometimes) shifting the equilibrium from a fully separating equilibrium to a semiseparating or pooling equilibrium, judicial review may reduce the amount of information that voters have about the incumbent leader’s type. One natural conjecture is that reducing information about the leader’s type tends to benefit the minority group and harm the majority group. After all, a voter might prefer a leader who shares her preferences because such a leader is likely to advance the voter’s interests on other, less easily observable issues. On this account, even when increasing the stringency of judicial review does not alter the minority group’s expected policy payoff (as when \( j \) decreases but remains in the \([pw, w]\) interval), this change may in fact benefit the minority (at the expense of the majority) by helping minority leaders remain in office, where they can work “behind the scenes” on the minority group’s behalf. If so, then judicial review indeed has countermajoritarian effects, but for reasons that are quite different from those usually emphasized.

There is, however, a plausible alternative conjecture: perhaps members of the minority group benefit from having leaders who overtly advocate for minority interests on high-profile issues, even if doing so makes those leaders more vulnerable to electoral defeat. Indeed, some scholars have suggested that it can be more valuable to minority groups to win some clear victories (even at the offsetting cost of more defeats) than to have a marginal influence on a larger number of issues (Gerken 2005a,b). If so, then countermajoritarian judicial review may *injure* the minority group’s true interests by inducing the pro-minority leader to pool with the pro-majority leader more often.

Second, judicial review in our model reduces the variance in policy outcomes. We have assumed that leaders are risk-neutral and so do not care about variance as such. However, if voters are risk-averse then judicial review would confer benefits on both the majority and the minority groups even when judicial review has no effect on the expected outcome. This potential benefit of judicial review—as a form of “insurance”—has been noted elsewhere in the literature (Hanssen 2004; Stephenson 2003). It is also possible, though, that the variance-reduction effect of judicial review is a cost for one or both groups, rather than a benefit. This
might be so if voters are risk-seeking, or if policy variation generates positive externalities, such as new information (Listokin 2008). Because our analysis indicates that judicial review will tend to reduce the variance in policy outcomes, a full assessment of judicial review’s welfare effects must take these considerations into account.

III. Uncertain Judicial Review

Part II assumed that if the leader proposes an illegal policy, \( x > j \), the court would always strike down that policy and replace it with \( x = j \). In other words, the preceding analysis presumed that the judiciary would consistently enforce a clear and well-known legal limit on government action. Indeed, a crucial assumption of the analysis in Part II was that a leader’s proposal of any \( x > j \) was welfare- and information-equivalent to a proposal of \( x = j \), which justified restricting attention to equilibria in which neither leader proposes any \( x > j \). The rationale for this assumption is that, if voters are savvy enough to recognize that when a legal limit is clear and consistently enforced, there is no meaningful difference between a leader who goes right up to that limit and a leader who goes beyond it (but who will certainly be pushed back). Thus in the model developed in Part II, judicial review never occurs in equilibrium, even though the existence of judicial review as an institution affects leaders’ behavior and voters’ welfare.

The assumption that the judiciary will consistently enforce a well-specified limit on anti-minority policy choices may make sense for some areas of law, where the minority-protective rules are sufficiently clear that no leader would ever violate them in equilibrium. In other areas of law, however, it is less clear whether or to what degree the judiciary will enforce legal protections for minority rights. This may be because of uncertainty about the legal views or policy preferences of the judges, because of uncertainty in the law itself, or because of uncertainty about whether the court will ever actually hear a legal challenge to the relevant policy. In this section, we investigate how the results may change if judicial invalidation of anti-minority policies is uncertain.
There are many ways one might model this sort of uncertainty. Here we explore a simple framework in which, as in Part II, there is a common knowledge threshold, \( j \), such that the court will always uphold any policy \( x \leq j \). In contrast to the earlier model, in which the court would always strike down any policy \( x > j \) (and replace it with \( x = j \)), we now assume that the court will strike down a policy proposal \( x > j \) with probability \( q \in (0, 1) \), where \( q \) is common knowledge. That is, if the leader proposes policy \( x > j \), then with probability \( q \) the final outcome is \( j \) and with probability \( 1 - q \) the final outcome is \( x \). One can interpret \( q \) as the probability that the court will conclude that the law in fact prohibits policies above \( j \), or as the probability that a court willing to enforce \( j \) as an upper limit in fact hears a challenge to the leader’s proposal.\(^{23}\) As in Part II, we will assume that voters can observe the leader’s proposal.

In this case, the utility to a leader of type \( \tau \) of proposing policy \( x > j \) is \( \hat{p}w - q|j - y_\tau| - (1 - q)|x - y_\tau| \). If the leader selects an \( x \leq j \), then her utility is \( \hat{p} - |x - y_\tau| \), as before. Let us consider first the pro-minority leader’s optimal strategy if we treat \( x_M \) as fixed and assume equilibrium voter beliefs. If \( x_M \leq j \), then Lemma 1 remains the applicable characterization of the pro-minority leader’s optimal strategy. If, however, \( x_M > j \), then the pro-minority leader’s optimal strategy is as follows:

**Lemma 3** When \( x_M > j \), and any \( x > j \) is struck down and replaced by \( x = j \) with probability \( q \), then:

(a) If \( x_M \geq \frac{w - qj}{1 - q} \), then the type-\( m \) leader selects her ideal point, \( x_m = 0 \).

(b) If \( x_M \leq \frac{pw - qj}{1 - q} \), then the type-\( m \) leader mimics the type-\( M \) leader by selecting \( x_m = x_M \).

(c) If \( x_M \in \left( \frac{pw - qj}{1 - q}, \frac{w - qj}{1 - q} \right) \), then the type-\( m \) leader plays a mixed strategy.

Now consider the incentives of the pro-majority leader. If the pro-majority leader selects her most-preferred policy \( (x_M = 1) \), then she receives her maximum policy payoff \( (0) \) if \( j \geq 1 \).

\(^{23}\)An alternative plausible formulation would make the probability of reversal an increasing and continuous function of \( x \). We defer that possibility to future research, and focus instead on the simpler setting in which the probability of reversal takes only two values \((0 \text{ or } q)\) depending on whether the proposal \( x \) is greater or less than a known threshold \( j \).
If \( j < 1 \), the pro-majority leader’s best option remains \( x_M = 1 \), but she now receives a lower policy payoff of \( q(j - 1) < 0 \). Therefore, if \( \frac{w-qj}{1-q} \leq 1 \), the pro-majority leader would always select \( x_M = 1 \), as this achieves perfect separation from the pro-minority leader (see part (a) of Lemma 3) without incurring any additional policy cost. However, if \( \frac{w-qj}{1-q} > 1 \), then the pro-majority leader is best off choosing \( x_M = \frac{w-qj}{1-q} \). The logic here is essentially identical to that underlying Lemma 2: it cannot be an equilibrium for the pro-majority leader to select an \( x_M \in [1, \frac{w-qj}{1-q}) \), because if the voters and the pro-minority leader expected her to play that strategy, and the pro-minority leader reacted accordingly using the strategies specified in Lemma 3, then the pro-majority leader could improve her payoff choosing a slightly higher \( x \), which would achieve perfect separation for only a trivial additional policy loss. We can state this formally as follows:

**Lemma 4** In equilibrium, the type-M leader chooses \( x_M = \max\{\frac{w-qj}{1-q}, 1\} \).

We can use Lemmas 3 and 4 to characterize the equilibria of the model with uncertain judicial review and observable policy choices as follows:\(^{24}\)

**Proposition 4** (a) If \( \max\{j, w\} < 1 \) and \( q \leq \frac{1-w}{1-j} \), then uncertain (countermajoritarian) judicial review induces a sincere separating equilibrium in which the type-M leader selects the popular policy \( (x_M = 1) \), which is upheld with probability \( 1 - q \) and replaced by \( x = j \) with probability \( q \) (such that the expected policy outcome under the type-M leader is \( 1 - q(1 + j) \)); the type-m leader selects the minority-preferred policy \( (x_m = 0) \), which is always upheld; and voters’ posterior beliefs are \( \hat{p} = 0 \) if \( x < 1 \) and \( \hat{p} = 1 \) if \( x \geq 1 \). The expected policy losses for each group are \( E(L_m) = E(x) = p(qj + (1 - q)) \)

\[ E(L_M) = 1 - p + pq(1 - j) \]

(b) If \( j \leq \min\{w, 1\} \) and \( q > \frac{1-w}{1-j} \), then uncertain (countermajoritarian) judicial review induces an extremist separating equilibrium in which the type-M leader selects \( x_M = \frac{w-qj}{1-q} > \max\{w, 1\} \), which is upheld with probability \( 1 - q \) and replaced by \( x = j \) with

\(^{24}\)As before, we restrict attention to cases in which judicial review may make a difference to the outcome, i.e. cases in which \( j < \max\{w, 1\} \).
probability $q$ (such that the expected policy outcome under the type-M leader is $w$); the type-m leader selects the minority-preferred policy ($x_m = 0$), which is always upheld; and voters’ posterior beliefs are $\hat{p} = 0$ if $x < \frac{w-q}{1-q}$ and $\hat{p} = 1$ if $x \geq \frac{w-q}{1-q}$. The expected policy losses for each group are $E(L_m) = E(x) = pw$ and $E(L_M) = 1 - p(2 - w) + 2pq(1 + j)$.

(c) If $w > j \geq 1$, then uncertain (anti-extremist) judicial review induces an extremist separating equilibrium in which the type-M leader selects $x_M = \frac{w-q}{1-q} > w$, which is upheld with probability $1 - q$ and replaced by $x = j$ with probability $q$ (such that the expected policy outcome under the type-M leader is $w$); the type-m leader selects the minority-preferred policy ($x_m = 0$), which is always upheld; and voters’ posterior beliefs are $\hat{p} = 0$ if $x < \frac{w-q}{1-q}$ and $\hat{p} = 1$ if $x \geq \frac{w-q}{1-q}$. The expected policy losses for each group are $E(L_m) = E(x) = pw$ and $E(L_M) = 1 - p(2 - w)$.

Proposition 4(a) establishes that so long as the leader’s reputational interest is not too strong (that is, so long as $w < 1$) and the probability of judicial reversal is not too high (that is, $q \leq \frac{1-w}{1-j}$), then introducing uncertain countermajoritarian judicial review has no effect on the leaders’ strategies. It does, however, improve the minority group’s expected payoff and reduce the majority group’s expected payoff. This result is intuitive, and essentially consistent with the results in Proposition 2 for equivalent parameter values.

Parts (b) and (c) of Proposition 4 reveal more subtle effects. Proposition 4(b) indicates that if judicial review is sufficiently countermajoritarian ($j \leq \min\{w, 1\}$) and the probability of judicial reversal is sufficiently high ($q > \frac{1-w}{1-j}$), then introducing uncertain judicial review will induce the pro-majority leader to propose a more extreme anti-minority policy than she would have in the absence of judicial review, while the pro-minority leader continues to propose her ideal point. Proposition 4(c) establishes a similar result for uncertain anti-
extremist judicial review. Such review induces the pro-majority leader to select a policy just extreme enough that the pro-minority leader would always propose her ideal point, and the policy that achieves this separation is even more extreme with uncertain judicial review than without any judicial review \((\frac{w-q}{1-q} > w)\). So, although introducing uncertain judicial review has no effect on the pro-minority leader’s policy choice (which remains \(x_m = 0\)), the pro-majority leader adopts a more extreme anti-minority proposal \((x_M = \frac{w-q}{1-q} > w)\). With probability \(q\), the court rejects this extreme policy (and replaces it with \(x = j\)), but with probability \(1 - q\) the court upholds (or fails to review) the policy.

This behavioral effect seems consistent with the idea that judicial review enables leaders to adopt more extreme positions than they would otherwise, because the court may rescue the leader from having to live with the consequences of the proposed policy (Salzberger 1993; Fox and Stephenson 2011). However, the dynamic here is different from the conventional version of this “bailout effect.” In our model, it is the pro-minority leader, not the pro-majority leader, whose strategy is directly affected by the prospect of a judicial bailout. The possibility of judicial reversal means that the pro-minority leader may not have to live with the consequences of a relatively extreme anti-minority policy choice. As a result, she is willing to go further in an anti-minority direction if doing so will fool the voters into thinking she is really pro-majority. In turn, the pro-majority leader must (and will) go even further in an anti-minority direction in order to achieve separation. In equilibrium, the pro-minority leader will not be willing to go that far, even with the prospect of judicial reversal, because there is always the possibility that the judiciary will not reverse the extreme proposal. Therefore, in equilibrium the pro-minority leader will continue to propose her ideal point, but the pro-majority leader—in the implicit shadow of the pro-minority leader’s threat to mimic—will adopt a more extreme anti-minority policy. Importantly, this does not occur when judicial reversal of any policy \(x > j\) is automatic (as in Part II), because in that case a proposal of \(x > j\) communicates no credible information to the voters. This further highlights a difference between our account and the more familiar “bailout” or “safety net” argument. On the conventional account, leaders are willing to make extreme proposals because they
(allegedly) know that the courts will strike them down, but (presumptively naïve) voters will still reward the leaders for making these proposals. In our account, it is the fact that the court might strike down an extreme proposal that makes such a proposal more attractive to the pro-minority leader, but it is the fact that the court might not strike down such a proposal that induces the pro-majority leader to become even more extreme, and the sophisticated (though incompletely informed) voters draw rational inferences from observed behavior.

Thus our model reveals that uncertain judicial review has two main effects on the policy choice, one direct and the other indirect. First, under minority-protective judicial review the courts will strike down some proportion of extreme anti-minority policies (the direct effect). Second, uncertain judicial review may induce the pro-majority leader to propose a more extreme anti-minority policy than she would in the absence of judicial review (indeed, even more extreme than the policy she would propose under certain judicial review). When the conditions described in Proposition 4(a) obtain, the indirect effect is not present, so judicial review unambiguously improves the expected policy payoff to the minority group at the expense of the majority group, and this distributive effect gets larger as \( j \) decreases (that is, as judicial review becomes more countermajoritarian). However, when the conditions in parts (b) or (c) of Proposition 4 obtain, the direct and indirect effects offset, such that uncertain judicial review has no effect on the minority group’s expected welfare.

Furthermore, if leaders have a strong enough interest in reputation (\( w > 1 \)), then uncertain judicial review can never improve the minority group’s expected welfare. When this reputational interest is weaker (\( w \leq 1 \)), then more minority-protective review (reductions in \( j \)) do improve the minority group’s expected welfare as \( j \) decreases from 1 to \( 1 - \frac{1-w}{q} \). This condition is derived by rearranging the bound on \( q \) in Proposition 4(a). Note that this bound is only constraining if \( q+w < 1 \); if that inequality is not satisfied, then \( 1 - \frac{1-w}{q} < 0 \), and \( j \) always falls above this threshold.
pro-minority leader would be willing to mimic this choice. (When $j$ is sufficiently low and $q$ is sufficiently high, the pro-minority leader views the proposal even of a fairly anti-minority policy as relatively low-cost, since the odds are that the final policy, after judicial review, will be close to the minority group’s ideal point.) When judicial reversal of any $x > j$ was certain, as in Part II, the minority group still benefitted (in expectation) from reductions in $j$. But when judicial review is uncertain, the minority group’s benefit may be offset by the fact that the pro-majority group makes an even more extreme proposal, and this proposal is sometimes upheld. So for some parameters there is a bound on the degree to which minority-protective (but uncertain) judicial review can benefit the minority group.\footnote{In particular, if $w + q < 1$, then the minimum expected policy loss to the majority group is $pw$; even setting $j = 0$ cannot reduce the minority group’s expected policy loss below this threshold.}

What about the welfare of the majority group? Parts (a) and (b) of Proposition 4 establish the intuitive result that the majority group’s expected policy payoff is lower with uncertain countermajoritarian judicial review than with no review. In the case described in Proposition 4(a), the direct effect of judicial review hurts the majority, and the indirect effect is inoperative. In the case described in Proposition 4(b), both the direct effect and the indirect of uncertain judicial review hurt the majority group: the pro-majority leader proposes a relatively extreme policy, which is sometimes reversed, meaning that from the majority group’s perspective, the final outcome even under a pro-majority leader is either too anti-minority (when the court upholds the proposal) or not anti-minority enough (when the court strikes down the proposal). Note that Proposition 4(b) describes a situation in which judicial review reduces the majority group’s policy payoff without any increase in the expected policy payoff for the minority group.

The preceding result is broadly consistent with the findings in Part II. However, we saw in Part II that when judicial invalidation of any $x > j$ is certain, anti-extremist review benefits the majority group without any reduction in the minority group’s expected policy payoff. This is not so when judicial review is uncertain. Proposition 4(c) shows that when judicial review is uncertain, anti-extremist judicial review has no effect on the majority group’s ex-
pected welfare (compared to the case with no judicial review). The reason is that in this case the direct and indirect effects of review offset: the possibility of judicial invalidation leads the pro-majority leader to make an even more extreme proposal, which is sometimes struck down (good for both groups) but sometimes upheld (bad for both groups). Moreover, when judicial review is uncertain, anti-extremist judicial review never induces the pro-minority leader to pool (fully or partially), because the pro-majority leader’s proposal is always extreme enough to induce separation. Thus when judicial review is uncertain, anti-extremist judicial review has no effect on either group’s expected policy payoff. Countermajoritarian judicial review may also have no effect if reputational interests are sufficiently strong, the probability of judicial reversal is sufficiently high, or if judicial review is too countermajoritarian.

When reputational interests are weaker and judicial review is less countermajoritarian, then countermajoritarian judicial review may benefit the minority at the majority’s expense, or may injure the majority without any expected benefit to the minority group.

Before proceeding, we make two additional observations regarding the welfare implications of Proposition 4. First, with uncertain judicial review, the voter can always distinguish pro-majority and pro-minority leaders, as was also true in the baseline no review case (Part I), but which was not true when judicial review was certain (Part II). Thus, in contrast to certain judicial review, uncertain judicial review does not affect welfare by altering the amount of information that voters have about the incumbent leader.

Second, the difference in the results for the case of certain and uncertain judicial review suggests that the impact of judicial review may change over time, as the scope and limits of the law become clearer. To illustrate, let us imagine (somewhat loosely and informally) a two-period variant on our model. In the first period, it is not entirely clear whether policies $x > j$ are unlawful; thus this first period corresponds to the “uncertain judicial review” model presented in Part III. However, the judiciary operates under a strong *stare decisis* rule, such that if the court rules at the end of the first period that a policy $x > j$ is unlawful, then in the second period all players know with certainty that any $x > j$ will be struck down (that is, the second period game will correspond to the “certain judicial review”
model developed in Part II). If the first period proposal is $x > j$ and the court upholds it, then in the second period all players will assume that there is no judicial limit on such proposals, and the second period game will therefore correspond to the baseline “no review” case presented in Part I. Finally, if the leader in the first period proposes $x < j$ (which will occur if the first-period leader is pro-minority), then the uncertainty remains unresolved, and the second-period game again corresponds to the model developed in Part III. Thus for the same underlying parameter values, and nominally the same judicial institutions, behavior may change over time as judicial rulings clarify (or fail to clarify) legal limits on anti-minority policies.

IV. Judicial Review of Unobservable Policy Choices

The preceding sections assumed that the leader’s choice of $x$ was directly and costlessly observable by the voters. While this is often a reasonable assumption, there may be situations in which voters have difficulty evaluating the effect of the leader’s policy choice on majority and minority interests. This could be because important aspects of the policy choice are literally secret, but the more common scenario is likely that the information necessary to map policy choices into distributional outcomes is difficult and costly for most voters to acquire and process. That is, we might suppose that the leader chooses a vector of policies, $V$, that maps into a distributional outcome, $x$, and that even if voters can observe $V$, they might be unable to infer $x$ (which is what they really care about). We therefore consider a variant on our model in which voters cannot observe the leader’s choice of $x$.\(^{32}\) We assume, however, that when the leader’s policy choice is subject to judicial review, the court can observe $x$, and the voters can observe the court’s decision whether to uphold or strike down the leader’s action. It turns out that this change in assumptions can lead to dramatically different conclusions about the effect of judicial review on the welfare of both the majority and the minority group.

\(^{32}\)The assumption of total unobservability, though extreme, simplifies the exposition.
As in Part II, we proceed by first deriving equilibrium behavior in the absence of judicial review, and then analyzing how and why the results may change in the presence of judicial review. When voters cannot observe $x$, equilibrium behavior in the absence of judicial review, described in the following proposition, is straightforward:

**Proposition 5** When voters cannot observe the policy choice $x$, there is a sincere equilibrium in which the type-$M$ leader selects the popular policy, $x = 1$, the type-$m$ leader selects the minority-preferred policy, $x = 0$, and voters’ posterior beliefs are $\hat{p} = p$. The expected policy losses for each group are $E(L_m) = E(x) = p$ and $E(L_M) = 1 - p$.

Now consider the effect of adding judicial review. As in Part II, we assume that judicial review prevents the leader from implementing any $x > j$, where $j$ is common knowledge. In Part II, we made the simplifying assumption that no leader would ever propose an $x > j$; we argued that this was a benign simplifying assumption because, in that model, any proposal of $x > j$ was information- and welfare-equivalent to a proposal of $x = j$. When $x$ is unobservable, this assumption is no longer innocuous, and so here, as in Part III, we allow the leader to select any $x$. As in Part II, however, any proposal of $x > j$ will be struck down, leading to a final outcome of $j$.

The interesting feature of this alternative model is that voters can use the court’s ruling to acquire information about the leader’s proposal (and the policy outcome). If the voters

\[33\text{For simplicity, we will assume here, as in Part II, that the court will always reverse any policy } x > j. \text{ We defer to future research the analysis of situations in which both judicial review is uncertain (as in Part III) and voters cannot directly observe policy (as in Part IV).}

\[34\text{The assumption that judicial reversal leads to a final outcome of } j \text{ is particularly important to the analysis in this section. This assumption means, in essence, that voters can observe when a leader pushes policy right up to the limit of what the courts will allow, but cannot distinguish among policy choices that stay well within that limit. This assumption is plausible in circumstances where judicial review has the effect of “trimming back” an unlawful distributional outcome—cutting out only those aspects of the policy that go too far—so that the final policy is just within the limits of legal acceptability. An alternative justification for our assumption would be that a conflict between the leader and the court—one that leads to judicial invalidation of the leader’s decision—generates publicity, and therefore greater attention to the leader’s (subsequent) choice (Flemming, Bolte and Wood 1997). If judicial invalidation simply led to the selection of a new $x$ by the leader, then judicial review would have no effect if } j \geq 1, \text{ and would shift the type-$M$ leader’s choice from } 1 \text{ to } j \text{ (while leaving the type-$m$ leader’s choice unchanged at 0) if } j < 1. \text{ On that alternative assumption, the results developed in this section would not hold. Another possibility is that if the leader proposes some } x > j, \text{ the policy reverts to some known status quo default, } q \in [0, j]. \text{ We show in the appendix that our qualitative results in this section hold in this alternative model, though the strength of those results is somewhat attenuated.} \]
observe a judicial reversal, they can infer that the leader initially chose some \( x > j \), and they also know that the final outcome was \( j \). If the court upholds the leader’s proposal, then the voters can infer that the leader implemented some \( x \leq j \), but they do not have any more precise information about \( x \). Thus in the case where voters do not directly observe \( x \), leaders may strategically provoke judicial invalidation in order to signal or conceal their type. The conditions under which they will do so are given by the following lemmas:

**Lemma 5**

(a) If \( j \geq 1 + (1 - p)w \), then both leader types choose their ideal point \((x_\tau = y_\tau)\).

(b) If \( j < 1 + (1 - p)w \), then the type-\( M \) leader provokes judicial reversal (and a final outcome of \( j \)) by selecting some \( x_M > j \).

**Lemma 6** Suppose the type-\( M \) leader provokes judicial reversal (and a final outcome of \( j \)) by selecting \( x_M > j \). The type-\( m \) leader’s best response, as a function of \( j \), is as follows:

(a) If \( j \geq w \), then the type-\( m \) leader selects her ideal point, \( x_m = 0 \).

(b) If \( j \in (pw, w) \), then the type-\( m \) leader plays a mixed strategy in which she selects \( x_m > j \) with probability \((p(\frac{w-j}{j}))\), and selects \( x_m = 0 \) otherwise.

(c) If \( j \leq pw \), then the type-\( m \) leader mimics the type-\( M \) leader by selecting \( x_m > j \).

The intuition for Lemma 5 is straightforward. If \( j > 1 \), the pro-majority leader can always guarantee herself a reputational payoff of \( pw \) if she chooses \( x_M = 1 \); if she provokes judicial reversal, thereby inducing a final outcome of \( j \), she incurs a policy loss of \((1 - j)\), but increases her reputational payoff by \((1 - p)w \).\(^{35}\) Her optimal choice depends on which of these is greater. As Lemma 5(a) notes, if \( j \) is sufficiently high that the pro-majority leader selects her ideal point, then the pro-minority leader also selects her own ideal point, and

---

\(^{35}\)These statements assume that the type-\( m \) leader always selects her ideal point. That assumption is reasonable. If the type-\( m \) leader selected an \( x_m > j \) with positive probability, it must be the case that her payoff from doing so is at least as high as that from selecting her ideal point. Because for all \( j > 1 \) the policy cost to the type-\( m \) leader of outcome \( j \) is greater than the policy cost to the type-\( M \) leader, but the reputational payoff is equivalent, it follows that whenever the type-\( m \) leader chooses an \( x > j \) with positive probability, the type-\( M \) leader does so with probability 1.
the outcome is the same as that without judicial review. If, however, judicial review has an effect—that is, if it induces the pro-majority leader to provoke judicial reversal—then the pro-minority leader’s optimal choice is given by Lemma 6. The intuition here, which closely tracks the intuition for Lemma 1, is as follows: The pro-minority leader must choose whether to select the minority group’s ideal policy or to mimic the pro-majority leader. The former is more attractive if the policy costs of mimicking are too high [part (a)]. When leaders care sufficiently about popularity, however, the pro-minority leader will mimic the pro-majority leader by provoking judicial invalidation sometimes [part (b)] or always [part (c)].

We can now use Lemmas 5 and 6 to characterize the impact of judicial review when voters cannot observe the policy choice \( x \), but can observe the judicial ruling. For expository clarity we will discuss separately the impact of anti-extremist judicial review \( (j \in [1, 1 + (1 - p)w)) \) and countermajoritarian judicial review \( (j \in [0, 1)).^{36} \)

The impact of anti-extremist judicial review, when the policy choice \( x \) is not directly observable, is given by the following proposition:

**Proposition 6**  
(a) Suppose \( w \leq 1 + (1 - p)w \) and \( j \in [\max\{w, 1\}, 1 + (1 - p)w] \). In this case, (anti-extremist) judicial review shifts the equilibrium to an extremist separating equilibrium in which the type-\( M \) leader deliberately provokes judicial invalidation by proposing some \( x_M > j \), leading to a final outcome of \( j > 1 \), while the type-\( m \) leader always selects \( x_m = 0 \), which is upheld by the court. Voters’ posterior beliefs are \( \hat{p} = 1 \) if the court strikes down the leader’s proposal, and \( \hat{p} = 0 \) otherwise. The expected policy losses for each group are \( E(L_m) = E(x) = pj \) and \( E(L_M) = 1 - p(2 - j) \).

(b) Suppose \( w > 1 \) and \( j \in (\max\{pw, 1\}, w) \). In this case, (anti-extremist) judicial review shifts the equilibrium to an extremist semiseparating equilibrium in which the type-\( M \) leader deliberately provokes judicial invalidation by proposing some \( x_M > j \), leading to a final outcome of \( j > 1 \), while the type-\( m \) leader selects \( x_m > j \) with probability \( \left( \frac{p}{1-p} \right) \left( \frac{w-j}{j} \right) \), and selects \( x_m = 0 \) otherwise. Voters’ posterior beliefs are \( \hat{p} = \frac{j}{w} \) if the

---

36Because judicial review has no effect when \( j \geq 1 + (1 - p)w \), we ignore that case.
court strikes down the leader’s proposal, and \( \hat{p} = 0 \) otherwise. The expected policy losses for each group are \( E(L_m) = E(x) = pw \) and \( E(L_M) = 1 - pw(\frac{j}{2} - 1) \).

(c) Suppose \( pw \geq 1 \) and \( j \in [1, pw] \). In this case, (anti-extremist) judicial review shifts the equilibrium to an extremist pooling equilibrium in which both leader types deliberately provoke judicial invalidation by proposing some \( x > j \), leading to a final outcome of \( j > 1 \). Voters’ posterior beliefs are \( \hat{p} = p \) if the court strikes down the leader’s proposal, and \( \hat{p} = 0 \) otherwise. The expected policy losses for each group are \( E(L_m) = E(x) = j \) and \( E(L_M) = j - 1 \).

Perhaps the most striking thing about the results stated in Proposition 6 is that anti-extremist judicial review, while ostensibly minority-protective, generates an expected policy outcome that is worse for the minority group than the expected outcome without judicial review. To see why, recall from Lemma 5(b) that when anti-extremist judicial review has any effect at all, it leads the pro-majority leader to implement a more extreme policy (\( j \) rather than 1). By picking a fight with the minority-protective court, the pro-majority leader reaps the reputational benefits of identifying herself as pro-majority. She could not accomplish this goal by selecting the majority group’s ideal policy because the voters cannot differentiate this choice from any other choice that is acceptable to the court. The pro-minority leader, meanwhile, either adopts exactly the same policy she would choose without judicial review, or else pushes a similarly extreme anti-minority policy.

The effect of anti-extremist judicial review on the majority group’s welfare is more ambiguous. For some parameter values, anti-extremist judicial review improves the majority group’s expected policy payoff, as was true when voters could observe \( x \) directly.\(^{37}\) However, for other parameter values, anti-extremist judicial review hurts both the minority and the majority.\(^{38}\) The reason is that, as noted above, anti-extremist judicial review has two effects.

\(^{37}\)For example, suppose \( p = 0.5 \), \( w = 3 \), and \( j = 1.1 \). In this case, the majority group’s expected payoff without judicial review is \(-0.5\), while the majority group’s payoff with judicial review is \(-0.1\).

\(^{38}\)For example, suppose \( p = 0.6 \), \( w = 1 \), and \( j = 1.2 \). In the absence of judicial review, both leader types choose their respective ideal policies, and the majority group’s expected policy payoff is \((0.4)(-1) + (0.6)(0) = -0.4\). Anti-extremist judicial review causes the pro-majority leader to pick an extreme policy, leading to a
First, as long as $j \in [1, w)$, judicial review may induce the pro-minority leader (sometimes) to implement $j$ rather than 0; this is bad for the minority group, but it is good for the majority group so long as $j < 2$. Second, anti-extremist judicial review induces the pro-majority leader to implement $j > 1$ rather than 1; this is bad for both the majority and the minority. For the majority group, these two effects may cut in opposite directions. From the majority’s perspective, although the greater transparency associated with anti-extremist judicial review redresses one agency problem (when the leader is pro-minority), it creates a second agency problem (when the leader is pro-majority). This result is thus consistent with other research showing that increasing transparency can sometimes exacerbate rather than ameliorate political agency slack (Fox 2007; Prat 2005).

Let us now consider countermajoritarian judicial review ($j < 1$). The results here are similar to those stated above for anti-extremist review, though the subsequent discussion highlights certain respects in which the welfare implications may be quite different.

**Proposition 7**  
(a) Suppose $w \leq 1$ and $j \in [w, 1]$. In this case, (countermajoritarian) judicial review shifts the equilibrium to a judicially-constrained separating equilibrium in which the type-$M$ leader deliberately provokes judicial invalidation by proposing some $x_M > j$, leading to a final outcome of $j \leq 1$, while the type-$m$ leader always selects $x_m = 0$, which is upheld by the court. Voters’ posterior beliefs are $\hat{p} = 1$ if the court strikes down the leader’s proposal, and $\hat{p} = 0$ otherwise. The expected policy losses for each group are $E(L_m) = E(x) = pj$ and $E(L_M) = 1 - pj$.

(b) Suppose $pw < 1$ and $j \in (pw, \min\{w, 1\})$. In this case, (countermajoritarian) judicial review shifts the equilibrium to a judicially-constrained semiseparating equilibrium in which the type-$M$ leader deliberately provokes judicial invalidation by proposing some $x_M > j$, leading to a final outcome of $j \leq 1$, while the type-$m$ leader selects $x_m = x_M$ with probability $(p - p)(w - j)/j$, and selects $x_m = 0$ otherwise. Voters’ posterior beliefs
are \( \hat{p} = \frac{j}{w} \) if the court strikes down the leader’s proposal, and \( \hat{p} = 0 \) otherwise. The expected policy losses for each group are \( E(L_m) = E(x) = pw \) and \( E(L_M) = 1 - pw \).

(c) Suppose \( j \leq \min\{pw, 1\} \). In this case, (countermajoritarian) judicial review shifts the equilibrium to a judicially-constrained pooling equilibrium in which both leader types deliberately provoke judicial invalidation by proposing some \( x > j \), leading to a final outcome of \( j \leq 1 \). Voters’ posterior beliefs are \( \hat{p} = p \) if the court strikes down the leader’s proposal, and \( \hat{p} = 0 \) otherwise. The expected policy losses for each group are \( E(L_m) = E(x) = j \) and \( E(L_M) = 1 - j \).

When judicial review is truly countermajoritarian, then the final policy outcome always lies between the ideal points of the majority and minority groups, which implies that, at least on the policy dimension, judicial review must either benefit one group at the other’s expense or else have no effect at all. That much is intuitive. What is more surprising is that although countermajoritarian judicial review sometimes benefits the minority group at the majority’s expense—as the conventional wisdom would predict—it is also possible for countermajoritarian judicial review to benefit the majority at the minority’s expense. To see why, recall that countermajoritarian judicial review has two effects. First, it constrains the pro-majority leader to implement \( j \) rather than 1, which is good for the minority and bad for the majority. Second, countermajoritarian judicial review may induce the pro-minority leader to implement \( j \) rather than 0, which is good for the majority and bad for the minority. This latter effect will outweigh the former effect if judicial review is not too countermajoritarian and leaders care sufficiently about popularity.\(^{39}\)

The discussion in this section has so far focused on the effect of judicial review on policy outcomes. Yet as we noted in Part II, the net welfare effects of judicial review may also depend on how it affects the information available to voters about the leader’s type, as well as the variance in policy outcomes. With respect to information, the results in Propositions 6 and 7 contrast sharply with the results in Propositions 2 and 3. When voters could

\(^{39}\)More precisely, countermajoritarian judicial review will, in expectation, benefit the majority at the expense of the minority when \( j \in (p, 1) \) and \( w > 1 \).
observe the policy choice $x$ directly and the court was sure to reverse any $x > j$, judicial review (weakly) decreased the amount of information available to voters, because judicial review sometimes induced a shift from a fully separating equilibrium to a semiseparating or pooling equilibrium. When voters cannot observe $x$, judicial review (weakly) increases the information available to voters by providing at least a partial indicator of the policy the leader chose. As Part II discussed, the normative implications of this result depend on whether greater separation tends to benefit the majority and hurt the minority, or whether the minority benefits more from separation than from pooling. Whatever one’s position on this question, the contrast between the results in Parts II and IV suggests that assessing the impact of judicial review on this informational aspect of voter welfare requires an evaluation of the extent to which voters can observe policy choice independent of the judiciary’s ruling. If, for example, greater type separation tends to benefit the majority at the expense of the minority, then (all else equal) judicial review tends to harm the minority and benefit the majority when policy is not directly observable, but tends to harm the majority and help the minority when policy is observable.

With respect to policy variance, recall from Part II that when voters can observe policy directly, judicial review (weakly) decreases the variance in policy outcomes. Propositions 6 and 7 show that the effects of judicial review on policy variance are less clear when voters cannot observe the leader’s policy choice. When $j$ is low (that is, under countermajoritarian review or relatively stringent anti-extremist review), judicial review will reduce policy variance, consistent with the conventional wisdom and the results in Part II. However, relatively lax anti-extremist review ($j$ close to $1 + (1 - p)w$) may increase policy variance by inducing the pro-majority leader to adopt a more extreme policy, while the pro-minority leader continues to adopt her ideal point much or all of the time.

The main insight derived from the analysis in Part IV is that because judicial approval or disapproval may provide more information to voters about the leader’s policy choice (and therefore her true preferences), the leader may have an incentive deliberately to provoke judicial reversal. Doing so credibly signals to voters that the leader has implemented the
most extreme policy that is legally permissible, \( j \), rather than any policy more favorable to the minority faction’s interests. Thus, precisely because the court is designed to protect minority interests (at least to some degree), the existence of judicial review creates incentives for politicians to pick fights with the court—pushing policy to legally-permissible limits—in order to demonstrate (or feign) allegiance to the political majority, rather than some unpopular minority.\(^{40}\) This can lead to perverse outcomes—in particular, the conclusion that, for certain plausible parameter values, ostensibly minority-protective judicial review reduces minority welfare and increases majority welfare.

**Conclusion**

Much of the extensive normative debate in constitutional theory over judicial review presumes that minority-protective judicial review leads to countermajoritarian outcomes—that is, outcomes that favor the protected minority at the expense of the majority. While our analysis partly corroborates this conventional wisdom, we also show that minority-protective judicial review may have other effects that qualify, and in some cases undermine, this hypothesis. These effects derive from the assumption that leaders whose preferences align with the majority have an incentive to reveal this fact, while leaders whose preferences align with an unpopular minority have an incentive to conceal this fact. Even when voters can observe the leader’s policy choices, the constraint imposed by minority-protective judicial review can induce a shift from a separating equilibrium (in which pro-majority leaders and pro-minority leaders cater to their respective constituencies) to a pooling equilibrium (in which both types of leader adopt the most anti-minority policy that the court would uphold). In other circumstances, the existence of a judicial “safety net” may make pro-minority leaders more

\(^{40}\)The extant literature has noted the phenomenon of elected officials deliberately enacting policies that the courts will reject; the usual explanation is that elected politicians can blame the courts for failure to enact policies favored by the public or by some powerful group (Graber 1993; Hirschl 2000; Salzberger 1993). Our model offers a related but distinct explanation. In contrast to the standard blame-shifting account, in our model leaders do not exploit voter confusion about which agents are responsible for policy outcomes, nor do voters necessarily prefer the judicially-invalidated policy to the lawful alternative (though they sometimes do). Rather, in our model leaders use conflict with the judiciary to signal their type.
inclined to propose relatively extreme anti-minority policies (which they secretly hope will be struck down) in order to mimic the pro-majority leaders, and this in turn induces pro-majority leaders to propose even more extreme anti-minority policies in order to distinguish themselves. Furthermore, if voters cannot observe the leader’s policy choices, then minority-protective judicial review has an additional effect: a leader who wishes to signal that she favors the majority’s interests (or to avoid signaling that she does not) has an incentive to provoke reversal by the court. Thus the introduction of judicial review may induce a shift from a world in which pro-minority and pro-majority leaders both choose their ideal policies to a world in which both of them implement the same policy—which may be adverse to the minority group’s interests, and perhaps even more extreme than the majority would prefer.

When these indirect effects are taken into account, one can no longer rely confidently on the intuitive assumption that minority-protective judicial review leads to countermajoritarian outcomes. In some cases, minority-protective judicial review may have no effect, or may benefit both the majority and the minority, or harm both the majority and the minority. Under some circumstances, judicial review may even benefit the majority at the expense of the minority. In such cases, majoritarian democrats should favor minority-protective judicial review, while those concerned with the welfare of unpopular minorities should oppose it.

Appendix

[TO BE COMPLETED]

References


