Fire Alarms and Democratic Accountability

Charles M. Cameron
Princeton University

Sanford C. Gordon
New York University

October 13, 2021

1 Introduction

Accountability exists in relationships between a principal and agent when the latter takes some action for which she may be made to answer by the former under a system of rewards and sanctions. In a democratic setting, the electorate are the principals and public officials the agents; rewards and sanctions the loss or retention of the benefits, privileges, and powers of office-holding. The hope is that in these settings, officials will have incentives to take actions consistent with the best interests of the citizenry; absent electoral accountability, citizens must fall back on the benevolence of their rulers or just plain luck. History shows these are weak reeds indeed. Thus, the accountability (or unaccountability) of elected officials to voters is a key component of any comprehensive theory of democratic governance.

Electoral accountability faces huge obstacles in practice. One of the most pernicious is the problem of asymmetric information: while knowing what politicians are up to is surely a critical ingredient of holding incumbent officeholders to account, for most voters, becoming and remaining informed about whether elected officials are actually meeting their obligations can be tedious, time-consuming and difficult. How many citizens have the time to research the voting record of their member of Congress, for instance? And even then, outside of a few specific areas, how many have the knowledge to assess whether a given legislative enactment actually improved their welfare, all things considered? A lack of transparency about actions and the obscurity between means and ends can render nominal accountability completely moot in practice.[1]

In the face of prohibitive information costs, a simple intuition is the following: *any mechanism that lowers informational costs for voters ought to enhance electoral accountability.* And, greater electoral accountability should make voters better off. Unfortunately, recent advances in the analysis of electoral accountability shows that this simple intuition can be wrong. For example, when politicians' actions are readily observable but their consequences delayed or obscure – the "wrong kind of transaprency" – politicians may have an incentive to "pander" – that is, to take

[1]To make matters worse, informed voters supply a public good – knowledgeable oversight of politicians – to uninformed citizens. If gathering information is costly for citizens, then members of the electorate face a collective action problem in which the uninformed can free-ride on the efforts of the informed. The resulting incentives may result in little information gathering and poor oversight of politicians [Hardin 2006]. But if information about politician behavior is effectively costless, then informational free-riding in the electorate becomes less concerning.

popular actions today even when they know full well the consequences may prove ineffectual or even harmful to voters ([Canes-Wrone et al 2001], [Maskin and Tirole 2004]; see also [Prat 2005], [Fox and van Weelden 2012]). Under such conditions, a degree of opacity concerning a politician's actions, if not the consequences of her choices, may soften the incentive to pander and actually improve the lot of voters. The lesson is that one must be very careful about overly simple intuitions when it comes to transparency and electoral accountability.

Nonetheless, the sense remains strong that if voters had access to mechanisms that lower their information costs — at least for the "right" kind of information — then democratic accountability might be enhanced, to the benefit of voters. What might those mechanisms be?

Fire Alarms

One important cost-lowering mechanism is *fire alarms*. In the context of accountability between a worker and a boss, a "fire alarm" connotes a readily perceived, reliable notice about bad worker performance that — critically — comes free or nearly free to the boss herself. In an electoral setting, a fire alarm consists of such a notice to voters about the performance of an incumbent. Might fire-alarms snatch electoral accountability from the jaws of information costs? That is our subject in this chapter.

Among political scientists, the idea of fire alarms gained currency from a celebrated analysis of congressional oversight of the bureaucracy ([McCubbins and Schwartz 1984]). In this setting, Congress (the boss or principal) has a hard time perceiving or evaluating the actions of bureaucrats (the workers or agent). McCubbins and Schwartz note that administrative procedures provide opportunities for private interests to relay information to Congress about bureaucratic noncompliance with legislative preferences. Relying on these actors — whom we will refer to generically as *sentinels* — is more efficient for Congress, they argue, than undertaking costly, active "police patrol" oversight. Thus, passive reliance on fire alarms need not imply congressional abdication of its oversight responsibility. Moreover, anticipation of the fire alarm may deter bureaucratic deviation from congressional desires. It seems plausible, then, that accountability can work well despite information costs, so long as some actor is willing to bear the informational costs for the boss.

Yet all may not be copacetic in the world of bureaucratic oversight. What happens when the sentinel is biased? Prendergast's ([Prendergast 2003]) game-theoretic analysis of service-providing

bureaucracies analyzes a strategic situation closely related to the one discussed by McCubbins and Schwartz, but in which the sentinel is not neutral in the signals it sends to a bureaucrat's political bosses. Much as students will only complain to their professors about unexpectedly low grades conferred by a teaching assistant but not unexpectedly high ones, so too will the bureaucrat's client only complain when a service or benefit is denied but not when one is granted – justifiably or not. Prendergast explores how this asymmetry affects the behavior of bureaucrats, and how their political principals may respond to neutralize the distortion. The overall lesson is clear: one needs to pay attention to sentinel bias.

Political scientists have explored the logic of fire-alarm oversight in other settings. A notable one is Supreme Court oversight of decisions in the U.S. Courts of Appeals. In this setting, the lower court (the "worker") is a three-judge panel deciding a case. The panel may do so in conformance with the preferred doctrine of the Supreme Court (the "boss"), but it may also deviate from the high court's preferred doctrine. For the high court, detecting such a deviation is difficult because, absent review, it is not privy to all the information available to the lower court. In an influential analysis, Cross and Tiller noted the following: if one of the judges on the panel is aligned with the higher court, she may announce a deviation via a dissenting vote ([Cross and Tiller 1998]). The dissenting vote from its ally is thus a fire alarm for the higher court, alerting it to the deviation and allowing it to review and correct the lower court's action. And, Cross and Tiller note, the possibility of fire-alarm dissents may in turn deter doctrinal deviations. Thus, fire-alarm oversight in the judicial hierarchy would seem to afford an easy path to judicial consistency and compliance with the rule of law.

Still, a caveat is once again in order. In a recent analysis, a group of scholars has returned to fire-alarm oversight in the judiciary, applying careful game theoretic reasoning to Cross and Tiller's informal arguments [Beim, Hirsch, and Kastellec 2014]. Their analysis confirms the intuitions of Cross and Tiller in some ways but also uncovers a potential issue: a "boy who cried wolf" problem. If the interests of the sentinel judge on the panel are too extreme, she may sound alarms too frequently, trying to induce the high court to review marginal deviations that it would prefer to let slide given the effort involved in reversing them. Crying wolf too often can lead the high court to ignore the dissenter's fire-alarm, which obviates the impact of fire-alarm oversight. Although this problem is different from those found in Prendergast's analysis of bureaucrats and their clients, the

lesson is similar: one needs to be very attentive to the interests of the sentinel.

The logic of fire alarm oversight would seem to extend naturally to an electoral context. Here, the seminal analysis was offered by Arnold ([Arnold 1993]). In an expansion of his argument in *The Logic of Congressional Action* [Arnold 1990], Arnold noted that fire alarms arise in the principal-agent relationship between voters and elected officials. He argued that political activists may have, and electoral challengers certainly do have, strong incentives to sound an alarm in the event of legislative malfeasance or error by a member of Congress. It is the threat of fire alarms, in this view, that might motivate legislator compliance with the preferences of inattentive citizens, just as the threat of fire alarms may induce bureaucrats to abide by the preferences of an inattentive Congress and lower court judges to attend to the doctrinal preferences of an over-burdened higher court.

It's worth reviewing Arnold's argument at length:

> The system contains activists who have incentives to monitor what legislators are doing in office and to inform citizens when legislators fail in their duties. Challengers to incumbent legislators have perhaps the strongest incentives for monitoring legislators' behavior and mobilizing voters. Few challengers fail to sift through incumbent voting records in search of issues that can be used against incumbent legislators. In addition, groups that bear major costs under a particular governmental policy may help publicize what incumbent legislators have done to contribute to their plight. Whereas challengers seek to replace incumbents, these groups may seek to persuade incumbents to avoid electoral repercussions by altering their positions and working for the groups' benefit ... Uncertainty abounds in a system like this. Legislators cannot possibly know for sure what policy effect will follow from specific governmental actions, how challengers or interest group leaders might use governmental actions or inactions to stir up citizens, or whether citizens might blame or absolve legislators for their connections to specific actions. What is certain is that legislators will do their best to anticipate citizens' preferences, to avoid the most dangerous mine fields, and to chart as safe a course as possible through the treacherous territory before them ... When legislators adjust their voting decisions to avoid generating preferences among inattentive citizens, is it fair

to suggest that legislators are controlled by those inattentive citizens? It is indeed. ([Arnold 1993]: 409, 411–12).

The critical point Arnold is making is that activists and challengers serve as sentinels, and that in this role their presence may enhance electoral accountability. As in the bureaucratic and judicial examples, the intuition is appealing. But the careful and subtle analyses of Prendergast in the bureaucratic setting and of Beim, Hirsch, and Kastellec in the judicial setting sound a warning: careful attention to sentinels' incentives is surely essential in mapping out the promise and perils of fire alarms for democratic accountability.

This Chapter

The aim of this chapter is to open the black-box of electoral fire alarms. We do not claim to offer more than a preliminary sounding of some rather deep waters.[2] Nonetheless we assay some provocative findings. In our analysis, the incentives of activists and those of challengers are quite different. As a result, fire-alarms from activists display many of the problems identified by Prendergast's analysis of bureaucratic fire-alarms. In contrast, fire-alarms from challengers have the potential to be extremely informative to voters, especially in tandem with information from the incumbents themselves. But there is an enormous caveat: the fire-alarm information must be nearly costlessly verifiable for voters. Absent "hard" information, challenger fire-alarms are vulnerable to attacks as "fake," "phony," or "fraud" — and often will be disregarded as a consequence. This caveat highlights the extreme importance of visible and believable sources of fact-checking, like trusted media outlets and investigative journalism. This caveat might have seemed of mostly theoretic interest when Arnold first raised the idea of electoral fire alarms in the early 1990s. But in today's American politics, its appears disturbingly relevant.

The chapter is laid out as follows. In the next section, we provide a simple review of modern electoral accountability theory as it bears on fire-alarm oversight of politicians. We are very selective as this literature has become enormous and forbiddingly complex. Yet in the area of electoral fire-alarms it is sparse. We then turn to a very simple model of electoral fire-alarms. Our presentation is non-technical, emphasizing the underlying intuition. The fourth section of the chapter offers an

[2]And, we are not the first. We review [Ashworth and Shotts 2011], the closest work, below. [Gordon and Huber 2007] contains an informal theoretical discussion that discusses the role of the challenger as fire alarm, and the incentives this creates for the incumbent.

application of some of the insights from our analysis to the politics of criminal justice. We conclude in the final section.

2 Sentinels and the Theory of Political Agency

The Voter and the Official

Our analysis of sentinels in the next section draws heavily on recent theoretical advances in the theory of political agency (TPA; see [Ashworth 2012] and [Gailmard 2014] for a review). Although our presentation is not especially technical, it may be tough going for those lacking familiarity with this branch of political economy. So, before tackling the analysis, we'll review a few of the basic ideas and building blocks in TPA, consider how sentinels fit into the standard account, and glance at related studies within the TPA tradition.[3]

For students approaching the study of accountability from the perspective of the mainstream political science research on the U.S., the first thing to understand is that TPA has distinct intellectual origins from classics in that literature by such scholars as David Mayhew, Richard Fenno, or R. Douglas Arnold. Likewise, if your understanding of the relationship between citizen and elected officials draws critical distinctions between prospective and retrospective voting; delegate and trustee models of representation; or substantive versus descriptive representation; you may find little familiar and much that is missing (at least at first glance).[4]

One important reason for the disjuncture is that those studies tend to focus on legislatures and legislators. A critical feature of legislatures, of course, is that they are bodies consisting of multiple agents. Consequently, much of the literature on accountability and representation in Congress focused on important features of this multiplicity: inter alia, the internal organization of the legislature, parties and party leadership, seniority, coalition building, and team production of legislation.[5] By contrast, scholars working in the TPA tradition — borrowing heavily from contract theory — have tended to abstract away from the multiplicity of legislators, focusing instead on the relationship between a single elected official (the incumbent) and an electorate that must decide periodically whether to retain her. Many models go even further, abstracting away from

[3] [Besley 2006] and [Fearon 1999] are clear and thoughtful introductions to TPA. [Duggan and Martinelli 2017] is rewarding for technically adept readers.

[4] For canonical research in these areas, see, e.g., [Pitkin 1967], [Key 1966], and [Fiorina 1981].

[5] For a recent example of an empirical investigation in this tradition, see Clinton et. al., this volume.

an electorate of many, heterogeneous voters in favor of a single, representative voter (for example, the median voter). As with all modeling enterprises, the point of this stark abstraction is not to deny the importance of the institutional and behavioral richness to be found in legislative settings. Rather, it is to hold fixed one set of relationships in the political environment in order to focus on and gain insight about another. This stripped down approach offers a clean way to consider some genuinely deep issues about elections and accountability, and accordingly, it is the approach we adopt in what follows. That being said, one could argue that the elected official envisioned by many models in the TPA tradition is an executive like a president, governor, or mayor, who can take some form of unilateral action that may be observed by the voters.

In the starkest models coming out of the TPA tradition, there are just two players: the incumbent official and the (representative) voter. A third player, the challenger, is often treated as a passive alternative, perhaps reflecting the intellectual origins of TPA approaches in contract theory, where replacing an agent entails a new draw from the labor market. Below, we will consider situations in which challengers are active players in their own right. An archetypal TPA game consists of two periods: in the first, an incumbent takes some action, which has some consequences for the voter; then, the voter may observe something — perhaps the action, perhaps the consequences (which one, as we will see, matters *a lot*) — and decides whether to retain the incumbent or replace her with a challenger. In the second period, the incumbent (or her replacement) again takes an action. This simple dynamic setting allows events and incentives to unfold in a natural way, and critically, introduces a motivation for voters to choose candidates they anticipate will act on their behalf in the second period.[6] The elected official cares about holding office, either because she receives direct benefits from holding office, or because she has policy preferences or abilities that diverge from those of her opponent, such that losing would be costly.

This archetypal framework contains the essential aspects of an accountability relationship between official and voter: the official, in her capacity as the agent of the voter, takes some action for which she may be evaluated and punished or rewarded by citizens through their vote choices.[7]

[6]Often, what will happen in the second period is so immediate that for convenience, the analyst will simply "roll up" the ensuing payoffs into first period payoffs and dispense with an explicit second period. We take this approach below, but the model is conceptually equivalent to one with two periods. One can, of course, extend beyond two periods, up to and including an infinite number.

[7]See Patashnik et. al. (this volume) for a discussion of potential biases in how voters evaluate

Given the sequence outlined above, the election dividing the two terms means that the voters casts ballots based on what they observe in the first period, with an eye toward what will happen (or is likely to happen) in the second. Hence, a critical insight of this literature is that the distinction between retrospective and prospective voting is really a distinction without a difference: voters form prospective forecasts of what *will* happen based on retrospective evaluations of what *has* happened.[8] Because what has happened is sunk cost, voting against an incumbent to punish her for misdeeds in office is not sequentially rational for a voter who thinks the alternative is worse. To be sure, in the real world there are surely voters that cast their votes out of pure emotion even if it means cutting off their nose to spite their face. The TPA approach sets such voters aside, not because its practitioners believe they don't exist, but rather with a specific analytical objective in mind: to isolate the incentives that may materialize for elected officials *even in the absence of any concern about vengeful voters* out to inflict punishment for past misdeeds.[9]

A critical feature of early TPA models (e.g., Ferejohn 1986) was the interchangeability of politicians (who all tended toward sloth or corruption). Voters' indifference among generic politicians allows the former to commit to a schedule of electoral rewards that induce effort on the part of the incumbent — for example, a promise to reelect a politician who surpasses a threshold level of performance. The implausibility of the incumbent homogeneity assumption driving accountability in these so-called "pure moral hazard" models (see [Fearon 1999] for an especially trenchant critique) has caused the vast majority of more recent research in the TPA tradition to take the heterogeneity of politicians as a starting point. In this view, elections are an institutional mechanism for differentiating "bad types" of politicians from good ones; the politician's actions (if observed) or their consequences (if observed) may provide useful information that facilitate the selection mechanism. A common feature of these more recent models is *hidden information*: the incumbent official may know her own type, or something about the state of the world about which the voter is comparatively ignorant.

information about incumbent performance.

[8] [Fiorina 1981] appears prescient about this TPA insight in way that, say, [Key 1966] was not. Fiorina envisions citizens who learn about the parties from their past performance, and then vote based on a prediction of what they will do in the future. So, retrospection leads to prospective-oriented voting. Key, in contrast, seems to envision citizens motivated by vengeance and gratitude, quite a different thing.

[9] If voters are driven by other emotion-based motivations, such as unreflective tribalism or nationalist frenzy, this will generally serve to weaken electoral incentives.

What makes an official "good" or "bad" depends on the context. For example, a good politician may be one with preferences or values that are *congruent* with those of the voters. Is an official the sort of person who would faithfully represent my interests in office *even in the absence of electoral pressures to do so?* As Bawn et. al. note in the chapter in this volume, sussing this out is a fundamental concern of organized interests weighing in on candidate selection in congressional elections. Examples of TPA models with preferences like this include [Morelli and Van Weelden 2013], [Wolton 2019], and [Snyder and Ting 2003]; the preferences in [Maskin and Tirole 2004] can be interpreted this way as well.

Good or bad could could also denote the politician's skill or expertise — does the official know what she's doing? Here, TPA models often employ a convenient device to explore its benefits: state-contingent preferences with private signals. In this approach, the voter wants the official to take an action appropriate to an imperfectly observed state of the world. A good example is counterterrorism: the policies we want our leaders to adopt in this area are highly contingent on whether the threat of a terror attack is imminent. We formalize this by having the agent receive a private signal about the true state of the world; skilled agents receive accurate signals while unskilled ones receive noisy ones (or perhaps nothing at all). In Section 3, we will adopt the technology of state-contingent preferences and type-dependent private signals. For simplicity we set preference congruence aside, leaving that as a topic for future research.

If politicians want to be reelected, then they will want to be perceived by the voters as good types. Will this lead those politicians to take actions that further voter interests? To answer this question, it is critical to consider what the voter can observe about an incumbent, and what inferences she can draw on the basis of what she has observed. An important distinction is between *actions* and *consequences*. In the counterterrorism example, this would be the distinction between an increase in police presence and airport security versus the occurrence or nonoccurrence of a terrorist attack. Table 1 displays a two-by-two contingency table arraying actions (observed or not observed) and consequences (observed or not observed). The result is four archetypal information environments. These are named following a taxonomy due to [Fox and van Weelden 2012].[10]

Under full transparency (FT), the voter can see both the incumbent's actions and the resulting

[10]Interestingly, the typology is analytically equivalent to Wilson's famous 2x2 typology of bureaucratic agencies ([Wilson 1989]).

Table 1: Archetypal Informational Environments in the Political Theory of Agency

Actions	**Outcomes/Consequences**	
	Observable	Not Observable
Observable	Full Transparency (FT)	Non-Transparent Consequences (NC)
Not Observable	Non-Transparent Actions (NA)	No Transparency (NT)

consequences. This situation would seem most favorable for the voter, both in terms of selecting good types and inducing the incumbent to act in the voter's best interests. However, when actions translate only probabilistically into outcomes, voters may still face a tricky inference problem – perhaps the official took an action that gave the best chances for the outcome preferred by the voter, but things went south by chance.

At the opposite extreme, under no transparency (NT) the voter can see neither actions nor consequences. Clearly, this is the situation most ripe for incentive and selection problems. As a practical matter, this environment is probably the most relevant baseline for most voters most of the time with respect to most policy making. Here, the voter will presumably have to rely on her presuppositions and presumptions about the incumbent and challenger.

An interesting environment is non-transparent action (NA) where the voter can observe outcomes but not actions. In our running example, suppose the voter observes no terror attack. Surely this is good news, but was there no attack because a secret counterterrorism operation was successful, or because there was no imminent threat to begin with? The NA case is applicable to "classic" retrospective voting with respect to economic performance: many factors affect unemployment, inflation, growth, and trade, perhaps more strongly than most actions a president can take. But if the economy is somewhat more likely to do better when the president takes the right actions and somewhat more likely to do worse when he takes the wrong ones, rewarding or punishing the president based on the economy's performance is sensible on the part of the voter, other things being equal.

Another well-studied environment is non-transparent consequences (NC). Here, the voter can observe the politician's action but cannot observe the consequences of the action, at least before the election. Famously, this situation can lead to "pandering," in which the politician takes a visible action the voter believes is correct — the popular action — even though the politician may know full well that it isn't ([Canes-Wrone et al 2001], [Maskin and Tirole 2004]). Suppose, for

example, an incumbent knows an attack is not imminent, but voters believe it is. That politician may implement costly but unnecessary policies to cater to the public's fear. Prat [Prat 2005] (see also [Fox and van Weelden 2012]) calls accountability based purely on observed actions "the wrong kind of accountability."

Sentinels as Third-Party Information Providers

We now have enough pieces in hand to consider how a sentinel may affect the electoral accountability game. A sentinel is a player who sees and reports 1) the politician's action, 2) the consequences of her action, or 3) the state of the world (when there are state-contingent preferences) at the time the incumbent took her action. A truthful report about these matters, if believed, has the effect of *moving the voter from one information environment to another*. So, using Table 1, a believable report about the official's action may shift the voter from NT to NC (e.g., "The President raised tariffs" or "The President abused his power.") It may shift the voter from NT to NA (e.g., "The deficit is simply enormous.") A believable report about the state of the world may shift the voter from NC to FT (e.g., "The President said he had to invade Iraq because of weapons of mass destruction, but it turns out there were no weapons of mass destruction.") Or a credible report about both actions and consequences may move the voter from NT to FT ("The president raised tariffs and the resulting trade war devastated farmers.") Depending on the ultimate information environment, the *threat* of the report may well induce an official to "do the right thing." But as we note above, it may also induce her to pander or otherwise take actions a fully informed voter might prefer she not take.

The point of a TPA analysis such as the one we conduct below is to consider how the introduction of a sentinel might affect accountability under different circumstances, and then evaluate the consequences for voter welfare. But before doing that, we need to answer two preliminary questions: what incentives motivate different sentinels to make reports, and what factors make sentinel reports credible for the voter? To get traction here, we distinguish three kinds of sentinels: *neutral conduits*, *interested parties*, and *challengers*.

A neutral conduit provides a handy analytic benchmark. Such a sentinel simply would report "the truth, the whole truth, and nothing but the truth" as it knows it. A neutral conduit is forthcoming and candid — if it knows something, it will report it truthfully, and cover up nothing.

And it is disinterested — it has no stake in the information itself or what the voter does with it. Neutral conduits approximate one ideal for the press, and for scientific experts. A strong intuition is, when neutral conduits exist, they will be very valuable to the voter.

In contrast, an interested party has a definite stake in a particular policy that it may hold *irrespective of the actual state of the world*, and potentially even irrespective of the identity of the incumbent officeholder. Suppose, for example, that the voter has state contingent preferences about infrastructure expenditures, of the following sort. The voter reasons, "If our national infrastructure is run down (State 1), then I favor a big infrastructure plan even if taxes have to go up. But if our national infrastructure is basically in decent shape (State 2) then I favor no plan and low taxes." So the voter wants the infrastructure plan matched to the state of the world. A neutral conduit would relay any available information to the voter about the state of the world that prevailed in the official's first term, and what the official did about it. But what will a concrete manufacturer or the association of civil engineers do? These actors are interested parties because they *always* favor a big infrastructure plan regardless of the actual state of our national infrastructure. So what can the voter infer from reports from civil engineers about the fitness of our national infrastructure? First, note that these parties will seldom pass on good news about the current state of infrastructure, news that would have motivated the incumbent to favor a small investment in infrastructure. The civil engineers' trade association will always give our bridges, roads, and airports a failing grade ([Society for Civil Engineers 2017]). This means that silence from an interested party may actually connote good news. Now suppose the voters receive the anticipated message from the interested party: "It's State 1 in America — our infrastructure's a wreck¡' What's a voter to believe? If the information in the report comes verified by a trustworthy source or is nearly costlessly verifiable by the voter himself — that is, if it is *hard information* — then the report is informative. But if the information is unverified and unverifiable, the voter should be skeptical. A report that always comes out the same way regardless of the true circumstances doesn't supply useful information about the state of the world; voters should disregard it. ([Milgrom 2008]).

Challengers are distinct from both neutral conduits and interested parties. As we discuss below, the key feature for the challenger is that he is in a zero-sum situation with the incumbent. Only one of the two contenders can win the election. Therefore, any information that hurts the incumbent is good for the challenger, and any that helps the challenger hurts the incumbent. Consider the

infrastructure example again, and assume that reports from the challenger are nearly costlessly verifiable. Suppose the incumbent has correctly matched the state, supporting the infrastructure plan if it was State 1 but opposing it if it was State 2. The challenger will not want to point this out to voters, as doing so will help the incumbent — it makes the incumbent look skillful. The challenger will therefore remain silent. However, if the incumbent failed to correctly match the state, the challenger will delight in pointing out the incumbent's blunder and apparent incompetence. The challenger's motives are distinct from those of the interested party because challengers are happy to make any report about the state of world, so long as it hurts the incumbent; and stay silent, regardless of the information available to them, when that information would help the incumbent.

3 Formalizing the Intuition

Preliminaries

To explore how the introduction of a sentinel can affect democratic performance – perhaps for the better, perhaps for the worse – we describe a highly stylized model. We begin with two players, an incumbent (she) and a voter (they), and then add a third: the sentinel (he). Drawing on the discussion in the previous section: the setting involves hidden information about the incumbent but in some scenarios also involves hidden actions as well; the model employs state contingent preferences; and the key hidden information about the incumbent is her skill in discerning the state of the world.

There is a state of the world, ω, which can take on values of 0 or 1. Both states are equally likely, and this is common knowledge. An incumbent can be described by her "type," t_i, which can be either low (L) or high (H). The incumbent initially knows her own type, but the voter has some uncertainty about the incumbent. Specifically, from the voter's perspective at the beginning of the game, the prior probability the incumbent is type H is given by α_i, and the probability that the incumbent is type L is $1 - \alpha_i$. The voter has a corresponding belief about the challenger's type, described by the parameter α_c. Both α_i and α_c lie between zero and one.

The incumbent receives a signal θ about the state of the world, which can take on values of either 0 or 1. The signal is correlated with the state of the world ω, but the quality of the signal depends on the incumbent's type. Specifically, high-type incumbents receive perfect signals about the state of the world ($\Pr(\theta = \omega | t_i = H) = 1$). In other words, if the state of the world is 1, a

high quality incumbent will know it's 1 for sure; if 0, she will know it's 0. Low-type incumbents, by contrast, receive imperfect signals. Specifically, their signals are correct with probability q (i.e., $\Pr(\theta = 1|\omega = 1, t_i = L) = \Pr(\theta = 0|\omega = 0, t_i = L) = q$), where q lies between $\frac{1}{2}$ and 1. We will often refer to q as the *quality* of the low-type incumbent's signal.

The incumbent must take a policy action $a \in \{0, 1\}$. In other words, she must choose between two policy actions, though one can interpret $a = 0$ as "do nothing" or "retain the status quo." The incumbent's action translates into consequences for voters in a simple way: if the incumbent's action matches the state, then the result is good for the voters. But if the incumbent fails to state-match, the result is bad for the voters. These payoffs are encapsulated in the following utility function:

$$u_v(a, \omega) = \begin{cases} 1 \text{ if } a = \omega \\ 0 \text{ otherwise} \end{cases}$$

In the interest of simplicity, we employ the following device. Let $\tilde{\alpha}_i(\mathcal{I})$ denote the posterior probability the voter assigns to the incumbent being high quality given information $\mathcal{I}$. We assume the probability the voter retains the incumbent is equal to $F(\tilde{\alpha}_i(\mathcal{I}) - \alpha_c)$, where $F(\cdot)$ is a smooth, strictly increasing function bounded between zero and one. This formulation is intended to capture the fact that while the incumbent benefits, *ceteris paribus*, from voters' positive impressions of her quality, other (stochastic) features of the political environment, as well as her impression of the challenger, may also affect the voter's ultimate selection of candidates. The key element in this formulation is the voter's posterior beliefs about the incumbent, which we assume are arrived at via Bayes' Rule wherever possible given available information.

We assume the incumbent wishes to maximize the probability she retains office. So, the incumbent (or her ally) will seek to maximize $\tilde{\alpha}_i(\mathcal{I})$, whereas an opponent (possibly a challenger, but also potentially a non-aligned media outlet) will seek to minimize it. The zero-sum nature of the competition between the incumbent and the challenger is a key point.

Setting up the model in this fashion "rolls up" play in the second period of an archetypal TPA game into the first period payoffs (a move noted in Section 2). In the one period game, in the absence of a sentinel, the incumbent faces an essentially decision-theoretic problem, though one strongly shaped by the voter's rational updating of beliefs. The addition of a sentinel then creates a game between the incumbent and the sentinel. Their strategic interactions shape voter posterior

beliefs and hence incumbent and challenger payoffs through the re-election function. This way of setting up the TPA game greatly simplifies the analysis and allows us to focus on the essential elements of fire-alarm accountability.

As discussed in Section 2, both incentive/sanctioning effects and politician selection effects are critical components of a theory of democratic accountability. Our model allows us to study both. A critical question we ask is the following: *can "virtuous behavior" by the incumbent benefiting the voter be sustained in equilibrium?*[11] In the context of our model, an incumbent who behaves virtuously is simply one who follows her signal. To see why, note that clearly, when the high-quality incumbent follows her signal, she assuredly state-matches since her high quality signal is correct with 100% probability. Hence, following the signal definitely leads to good consequences for the voter. But since we assumed both states of the world are equally likely to begin with, and, further, that the low-quality incumbent's signal is right more often than not (even if not especially often), when the low-quality incumbent follows her signal she maximizes the expected benefit to the voter as well.

The second critical component in the theory of democratic accountability concerns how much voters learn about the incumbent's quality, thus permitting them to make more informed choices at election time. So the issue is, in equilibrium do voters learn much about the incumbent's type? While our simple model has only one period, the probabilistic vote function effectively captures a more complicated, multi-period model in which voters enjoy downstream benefits from having a high-type incumbent in office. These benefits are more likely to accrue to voters, *ceteris paribus*, when they have better information about the incumbent's type, as it will permit them to make fewer errors in determining whether the incumbent or challenger is the better choice.

We now turn to the four information environments in which voters might find themselves *in the absence of a sentinel*. The four information environments were indicated in Table 1.

Voter observes neither policy action nor outcome. This is the NT (No Transparency) scenario in Table 1. This information environment is the simplest case. If the voter can observe neither the incumbent's action a nor whether it was correct (i.e., whether $a = \omega$), there is nothing that the incumbent can do to alter the voter's beliefs. The voter will then reelect with probability

[11] This does not necessarily imply or require that virtuous behavior be the unique equilibrium. For example, in one of our baseline examples, any behavior by the incumbent, including virtuous behavior, can be sustained as an equilibrium.

$F(\alpha_i - \alpha_c)$. Note that because the incumbent's electoral fortunes are unaffected by her policy choice, any policy choice in any incumbent information set is an equilibrium. Of course, this includes "doing the right thing" by following the signal. We can easily break this indifference by giving the incumbent some infinitesimal benefit from pursuing the voter's interests. In that case, behaving virtuously is the unique equilibrium. But voters still won't learn about the incumbent's type from that behavior.

Voter observes policy action, but not outcome. This is the NC (No Consequences) scenario in Table 1. Such an information environment is almost as simple as the NT baseline. To see why, suppose that in equilibrium, both types of incumbent behave virtuously. Because we have deliberately set up the model so that both states of the world are equally likely, and so that the accuracy of the incumbent's information is independent of the state, the voter will be just as likely to observe $a = 0$ as $a = 1$. Hence, the policy will provide no new information about the incumbent's type. And given this, the incumbent has no incentive to deviate from virtuous behavior – just like in the case in which the voter observes neither the policy action nor the outcome.

Note that this virtuous behavior is sustainable because of a deliberate modeling choice on our part — making both states of the world equally likely. This eliminates any incentive of the incumbent to pander to voters by choosing the policy more consistent with their prior belief.

Voter observes the outcome, but not the policy action. This is the NA (No Actions) scenario in Table 1. In this informational environment, the voter learns whether the incumbent was right or wrong in their policy choice, even though she cannot observe incumbent actions. Now suppose both types of incumbent behave virtuously. High-quality incumbents will always choose the correct policy, and low-quality incumbents will choose the correct policy more often than not. Given that the incumbent behaves virtuously, the voter, upon learning that the wrong policy was chosen, will know with certainty that the incumbent is low quality ($\tilde{\alpha}_i(a \neq \omega) = 0$), thus decreasing the probability the incumbent is retained. If, by contrast, the voter learns that the correct policy was chosen, then the voter will know that it was chosen either by the high quality incumbent or by a low quality incumbent who received a signal that turned out to be correct. The voter's posterior belief on the incumbent will then (by Bayes' Rule) be equal to, $\tilde{\alpha}_i(a = \omega) = \frac{\alpha_i}{\alpha_i + (1-\alpha_i)q}$, which is strictly greater than α_i, thus increasing the probability the incumbent is retained. Given the foregoing, the low-quality incumbent has every incentive to get the policy right, She maximizes the

Table 2: Summary of Baseline Cases Absent a Sentinel

Actions	**Outcomes/Consequences**	
	Observable	Not Observable
Observable	FT: virtuous equilibrium unique, some updating	NC: virtuous equilibrium possible, no updating
Not Observable	NA: virtuous equilibrium unique, some updating	NT: virtuous equilibrium possible, no updating

probability of getting the policy right by behaving virtuously.

Voter observes both the outcome and the policy action. This is the FT (Full Transparency) scenario in Table 1. Given the setup of the model, the logic is identical to the previous case.

We summarize the no-sentinel baselines in Table 2, which mirrors Table 1. It is worth noting, in terms of motivating virtuous actions, our simplifying assumptions bias the model in favor of the incumbent doing right by the voter. This allows a clear baseline from which to consider whether the strategic game between the incumbent and the sentinel actually improves voter welfare.

Introducing the Sentinel

Our sentinel-free analysis demonstrates how access to verifiable information about the outcome can enhance accountability, both by strengthening the incentives for the incumbent to behave virtuously, and by enhancing the ability of voters to select good types. We also see the limitations of informational environments in which the voter can observe the incumbent's action but not the consequences of the choice.

With these considerations in mind, consider the following permutation of the model. Suppose we are in a world in which absent any fire alarm, the voter observes neither the policy nor the outcome (the NT information environment). A third party, whom we call the sentinel, s, receives some information that he may pass on to the voter. Specifically, if the incumbent chooses policy action $a = 1$, the sentinel receives evidence that the policy was either right (if $\omega = 1$) or wrong (if $\omega = 0$) with probability π_1; with probability $1 - \pi_1$, the sentinel receives no such information. Likewise, if the incumbent chooses policy action $a = 0$, the sentinel receives evidence that the policy was either right or wrong with probability π_0, and with probability $1 - \pi_0$ receives no such information. Note that π_1 need not equal π_0.[12]

[12]This setup resembles the "asymmetric resolution" extension in [Canes-Wrone et al 2001].

There are now several things to consider. First, consider the preferences of the sentinel, as discussed in Section 2:

- The sentinel may be a *neutral conduit*, sharing the voter's preference for a high quality incumbent;
- The sentinel may be a *challenger*, who benefits when the incumbent's reputation suffers; or
- The sentinel may be an *interested party* that is biased in favor of one policy (say, $a = 0$), irrespective of the actual state of the world.

Obviously, in a world of self-interested political actors, we would have strong reason to believe that the sentinel is unlikely to be a neutral conduit. Nonetheless, as noted above, a neutral and honest sentinel is a useful benchmark against which to compare the challenger and activist sentinels.

Next, consider the nature of the evidence the sentinel may pass on to the voter. Two extreme cases are the following:

- The information may be *cheap talk*, in the sense that there is nothing to validate the veracity of the information other than the voter's belief that the sender is being honest; or
- the information may be *hard*, in the sense that it will be taken as fact by a voter presented with it. (For example, the message may be costlessly verifiable or falsifiable.)

Of course, many other kinds of evidence are possible. Information may be verifiable at cost, or verifiable at cost with some probability, for example. To convey the intuition as expeditiously as possible, we will abstract away from these cases and consider only the more stark examples above.

3.1 The Cheap Talk Fire Alarm

Our first step in demonstrating the contingency of the purported democracy-enhancing properties of fire alarms is to consider them when evidence is cheap talk. Under such conditions, for informative communication between the sentinel and the voter to be possible, the latter must take the former's word as given.

Neutral Sentinels with Cheap Talk. First, consider the neutral sentinel, who is known to be neutral by the voter. By construction, this actor shares the preferences of the voter, and is thus, in a sense the perfect agent. Because of the common interest in high-quality incumbents, an

equilibrium exists in which the sentinel neutrally and honestly conveys information to the voter (when he has it) and the voter believes the sentinel and acts accordingly.

Moreover, consider the incentive effects of the neutral sentinel on the reelection-minded incumbent. Recall that high-quality incumbents can always choose the policy that is correct from the voter's perspective. Suppose she does so. Then the sentinel will either have good news to pass on to the voter (the policy was correct), or no news. By contrast, the low quality incumbent may err, generating bad news that the sentinel will dutifully, and credibly, pass on. Bad news perfectly reveals that the incumbent is a low-type. Clearly, then, low-quality incumbents have an incentive to minimize the probability of bad news. They do this by behaving virtuously.

If we understand "fire alarms" to be signals purporting malfeasance (bad news), then in the presence of a neutral sentinel, there is an equilibrium in which fire alarms (1) are taken seriously by voters; (2) strengthen the incentives of the incumbent to behave virtuously; and (3) help voters make more informed choices come election time. Note that this intuition would be preserved even if the only evidence available to the neutral sentinel were evidence of bad news – i.e., the original conception of fire alarms per McCubbins and Schwartz and, later, Arnold. Incumbents would still be motivated to minimize the probability of bad news, which they accomplish by behaving virtuously. This situation seems to confirm Arnold's optimism about sentinels in an electoral context.

Challenger Sentinels with Cheap Talk. Of course, in politics, neutral conduits may be hard to come by. As noted in the Arnold quotation in the Introduction, the sentinel is much more likely to be a self-interested actor like a challenger or activist, whose interests may not perfectly align with those of the voter. It turns out that if talk is cheap, there exists *no informative equilibrium* in which the voter takes statements by the sentinel as credible.

It is easy to see why. If the sentinel is a challenger, his objective is to damage the incumbent's reputation. Now suppose, again, that high-quality incumbents always choose the policy that matches the state of the world. Irrespective of the information actually received by the sentinel, he will always have an incentive to try to convince the voter that he received information about incumbent performance and that it was bad – i.e., that the policy chosen did not match the state. But because this motivation persists regardless of the truth, the voter *will learn nothing* from his utterances. Not all is lost, however: the incumbent will still weakly prefer to pursue the policy expected to most benefit the voter, just as she did in the case with no sentinel.

Interested-Party Sentinels with Cheap Talk. Sometimes the sentinel is an activist or advocacy group. Are things better when the sentinel is an interested party? Not really. Unlike the voter (and the challenger), the activist does not care about the occupant of the office. He is happy as long as his favored policy is pursued. Consequently, any pronouncements he makes will be independent of the actual state of the world. If the incumbent chose the favored policy, he may announce that this was the correct; likewise, if the incumbent chose the disfavored policy, he may announce that it was wrong. But because he will never denounce the favored policy choice or praise the disfavored one, the voter will dismiss those other utterances as not credible, and learn nothing from them. But also again, the incumbent will continue to at least weakly prefer to do right by the voter.

The upshot of the foregoing is that if information is cheap talk, the possibility that fire alarms will improve the voter's lot compared to a counterfactual world with no fire alarms is contingent on very specific assumptions about the preferences of the sentinel pulling the alarm.

3.2 Fire Alarms with Hard Evidence

A reader might, given the foregoing, be tempted to jump to one of two conclusions: either fire alarms only benefit the voter under rare circumstances (the truly neutral sentinel), or the failure of fire-alarms to help voters is a mere artifact of the artificial cheap talk environment in which we have considered them. This conclusion would be premature at this point, however. We first need to consider how things work in a different informational environment. Here we do so by supposing that information is hard rather than cheap talk. In other words, we consider a setting in which, if the sentinel receives "good" news that the policy correctly matched the state, that evidence is both accurate and dispositive: thus, if the sentinel provides the news to the voter, the latter will accept it at face value. Likewise, if the sentinel has "bad" news and transmits it to the voter, the voter will accept it as true. In such a situation, we will also need to consider what voters will believe if they receive *no* news. No news admits two possibilities: either the sentinel had no information to transmit; or, the sentinel had information but suppressed it.

Neutral Sentinels with Hard Information. In this stark informational environment, consider first the neutral-conduit sentinel. In the cheap talk setting, this actor conveys his possession of good news, bad news, or no news, and the credibility of these signals is established endogenously

in equilibrium given the likemindedness of the voter and sentinel. In other words, "the truth will out," at least to the extent that there is truth available to the sentinel to reveal. Moving away from cheap talk to a hard information environment changes nothing: the sentinel will still report information available to him, and now, the evidence will be even firmer than it was in a situation where it was already accepted at face value. Consequently, the truth will continue to out. Voters will have access to all available information, facilitating their goal of selecting the best available officeholder. And incumbents, for their part, will have strong incentives to behave virtuously, as doing so maximizes their probability of reelection.

Challenger Sentinels with Hard Information. In a world where high quality incumbents always choose the correct policy, the challenger has a clear incentive to report all bad news regarding the incumbent's choice to the voter, because the voter will infer from bad news that the incumbent was a low type, thus enhancing the challenger's electoral prospects. Just as bad news undermines the incumbent's reputation in the eyes of the voter, so too does good news enhance it; hence, the challenger will suppress any good news he has at his disposal. If a challenger sentinel reports all bad news and suppresses all good news, and voters interpret bad news as definitively establishing the incumbent as a low-type, then a high-type incumbent will clearly have an interest in selecting the correct policy: doing so guarantees that no news will be conveyed to the voter, whereas deviating and choosing the wrong policy will result in a lottery between the electoral consequences of no news and bad news. No news is better.

But now consider the problem from the perspective of the low quality incumbent: she has an incentive to minimize the likelihood of bad news. This was also the case when the low quality incumbent faced a neutral sentinel, of course. But given the special incentives of the challenger, does this situation now mean a departure from virtuous behavior by the incumbent? It turns out that the answer is *yes*, at least under some conditions.

Specifically, suppose $\pi_1 > \pi_0$, so that the challenger is more likely to receive hard information to pass on to voters when the incumbent has chosen $a = 1$ (e.g., depart from the status quo) than when she has chosen $a = 0$ (e.g., maintain the status quo). If the incumbent is a low type, and receives a signal of $\theta = 0$, she has two reasons to follow her signal and choose the corresponding action of $a = 0$, to the benefit of the voter: the choice is more likely to be correct, and if it is incorrect, it is less likely to yield bad evidence for the challenger to transmit to voters.

If the incumbent is low quality and receives a signal of $\theta = 1$, however, she faces a tradeoff given the threat of a fire alarm: if she chooses $a = 1$, she is more likely to be correct, but if she is incorrect, the challenger will "have the goods" on her with relatively high probability. If, on the other hand, she chooses $a = 0$, she is less likely to be correct, but if she chose incorrectly, it is less likely that the challenger will actually receive the bad news to pass on to the voters. It turns out that given a signal $\theta = 1$, the low quality incumbent will take the non-virtuous action that *hurts* the voter in expectation ($a = 0$) if and only if

$$\frac{\pi_1}{\pi_0} > \frac{q}{1-q}.$$

In other words, if the risk of bad news associated with choosing the policy that is correct in expectation is sufficiently great relative to the low quality incumbent's expertise, the low-quality incumbent will choose the policy more likely than not to be *wrong.*

The upshot of the foregoing analysis is that the presence of the sentinel threatens to create a distortion in the incentives of (low-type) incumbents relative to a world with no sentinel. Note, however, that the effect of the sentinel's presence on the voter's overall well-being is ambiguous. This is because relative to the baseline with no sentinel, voters can update their beliefs about the incumbent's type: those beliefs will be revised downward (to zero, in fact) given the provision of bad news; and they will be revised upward given the provision of no news. Hence, our partially-informed voter is more likely to get a high-type incumbent in office after the election. The incentive and selection effects from a challenger sentinel are in tension, making overall conclusions regarding voter welfare ambiguous in the absence of stronger assumptions.

Interested-Party Sentinels with Hard Information. Next, suppose the sentinel is an interested party – say an activist or advocacy group – and in particular, one who cares only that the incumbent selects the policy action $a = 0$, irrespective of the state of the world. With these preferences, the group is indifferent with respect to the occupant of the office, and can craft a fire alarm strategy that maximizes the incumbent's incentive to choose $a = 0$, even when she receives the signal suggesting she ought take the opposite course ($\theta = 1$). Consider the following strategy for the sentinel: report available good news if and only if the incumbent turns out to have correctly chosen the action $a = 0$; and report available bad news if and only if the incumbent turns out to

have incorrectly chosen $a = 1$. Given this strategy from the sentinel, a high quality incumbent who always follows her signal need never fear bad news – from her perspective, no news is the worst possible outcome. And so, the voter could again infer from bad news that the incumbent was low quality with certainty.

Now consider the game from the perspective of the low quality incumbent. She will, of course, prefer to follow her signal given $\theta = 0$ – doing so maximizes the odds that the voter will receive good news about her performance. The interesting question concerns what she will do when $\theta = 1$, which could place her at odds with the interested-party sentinel. Suppose she behaved virtuously in that circumstance and chose $a = 1$. Then with probability $(1 - q)\pi_1$, she would be wrong, and the advocacy group would learn about it and publicize the bad news to the voter, driving the incumbent's reputation $\tilde{\alpha}_i$ down to zero. With complementary probability $(1 - (1 - q)\pi_1)$, the voter would receive no news. But recall that from the voter's perspective, receiving no news is also consistent with the incumbent being high quality. Clearly, the low quality incumbent would prefer no news to bad news.

Given this preference, the low quality incumbent can improve her lot by deviating from virtuous behavior by choosing policy $a = 0$ when her signal is $\theta = 1$. With probability $(1 - q)\pi_0$, this was actually the right move and the advocacy group will have good news to convey to the voter about the incumbent's correct choice. With complementary probability, there will either be no news to report, or bad news that the interested-party sentinel will suppress – hence, no news from the perspective of the voter. But a lottery between good news and no news is clearly better from the low-quality incumbent's perspective than a lottery between bad news and no news. Hence, virtuous behavior by the low quality incumbent given $\theta = 1$ cannot be an equilibrium. Given reasonable restrictions on beliefs off the path of play (namely, if we assume the voter will infer the incumbent is low quality given bad news), we can establish the converse: the low quality incumbent's disregarding the signal $\theta = 1$ *is* consistent with equilibrium play. As in the case of the challenger sentinel, the activist sentinel induces incumbents sometimes to do the wrong thing from the voter's perspective. Note that unlike in the case of the challenger sentinel, this will be the case irrespective of the underlying parameters (q, π_0, and π_1).

Despite the bias in reporting from an interested-party sentinel, voters will, on occasion, have access to hard information about the incumbent's performance (although, in equilibrium, the voter

will never receive any bad news due to the distortion in the low quality incumbent's behavior). Accordingly, the voter's selection problem will be mitigated relative to a baseline in which the voter has no access to information about policy or performance. As in the case with the challenger sentinel, however, there is again a trade-off from the voter's perspective between the selection benefits and the incentives for some incumbents to do the wrong thing, some of the time.

3.3 Can Incumbent Credit-Claiming Mitigate Biased Fire Alarms?

Up to this point, we have considered cases in which the sentinel is the only source of information available to the voter. Of course, we know that incumbents are generally eager to claim credit for any good news, whether or not it is actually associated with their actions in office ([Mayhew 1974]). A natural question to ask, then, is whether extending the model to permit the incumbent to transmit good news to the voter about their performance might mitigate some of the distortions associated with biased sentinels that we described in the previous section. Needless to say, if communication is pure cheap talk, there can be no credible communication between the re-election-minded incumbent and the voter: the former will always want to convince the latter she is a high-type, irrespective of her actual type.

Accordingly, we restrict our attention in this section to the hard information environment, and assume for simplicity that if there is information to pass on to the voters, it will be in the possession of both the incumbent and the sentinel . To the extent that "the truth will out" given a neutral sentinel, the information conveyed to the voter via credit-claiming will be redundant. Accordingly we would anticipate no changes to the welfare of the voter – either through the selection or incentive channels – given the presence of a credit-claiming incumbent and neutral sentinel.

Consider, next, the case of the challenger sentinel. Recall that the challenger transmits information to the voter if and only if it is bad news for the incumbent. By the same logic, the incumbent will only transmit information to the voter if it is good news (remember – if high quality incumbents have the ability to select the correct policy with no error, then bad news will fully reveal that the incumbent is low quality). So with the addition of a credit-claiming incumbent and a detractor challenger, *all available evidence will ultimately be passed on to the voter*: good news by the incumbent, and bad news by the challenger. But then, from the perspective of the voter, we are in a world that is equivalent to one with the neutral sentinel, which we have already estab-

lished is excellent from the voter's perspective: the selection problem is mitigated because of the rich information available to the voter; and there are no distortions in the low quality incumbent's incentives. Interestingly, this situation closely resembles an idealized view of an adversarial legal system, in which prosecutors have incentives to present all available incriminating evidence to a jury, while defendants have incentives to present all available exculpatory evidence (Dewatripont and Tirole 1999).

The question of whether credit-claiming by the incumbent can mitigate the biased information coming from an interested-party sentinel is more subtle. To be sure, when we allow for credit-claiming with hard information, more information can potentially reach the voter. Specifically, suppose a low-type incumbent were to choose the policy disfavored by the activist. In the absence of credit-claiming, the voter would only observe no news or bad news. Add a credit-claiming incumbent and now good news is also a possibility. More information will be revealed than before, and the downside electoral risk of crossing the interested-party sentinel will be muted.

Suppose, by contrast, the incumbent chooses the policy the activist wants despite receiving the signal suggesting the contrary action. Now, both the incumbent and the activist have a mutual interest in suppressing any bad news about that policy. By catering to the activist's preferences to the detriment of the voter, the worst outcome for the incumbent – bad news – can be avoided. Implicitly, the incumbent colludes with the interested party, against the voters.

The downside risk of bad news to the low quality incumbent may thus continue to motivate her to insure against its adverse consequences by selecting the policy favored by the interested-party sentinel, even if she believes it is the wrong policy. However, this downside risk is offset to some extent by the potential upside: credit-claiming if she chooses the policy the advocacy group dislikes and it turns out to have been the right choice ("standing up to the special interests").

Given the foregoing analysis, introducing hard information credit-claiming doesn't bring us quite back to a situation equivalent to that of the neutral sentinel, as it did in the case of the challenger sentinel. In supplementary analysis, we demonstrate the existence of conditions under which virtuous behavior cannot be sustained in equilibrium, even given the otherwise ameliorating effects of credit-claiming. In other words, the distorting effects of a biased sentinel cannot be fully eliminated when the sentinel is an interested party. Here are some observations concerning the conditions under which the incumbent will succumb to the activist's demands (choosing $a = 0$ even

given a signal of $\theta = 1$) to the detriment of the voter:

First, unsurprisingly, the low quality incumbent will be more inclined to implicitly collude when q, the accuracy of her own information, is relatively low, because the risk that the interested-party sentinel will publicize bad news if she sticks with her signal and chooses $a = 1$ will be correspondingly high.

Second, consider that in previous cases, the incumbent's choice was relatively simple: she was, for example, choosing between lotteries of no news and bad news, or comparing a lottery between no news and good news with a lottery between no news and bad news. By contrast, with an interested-party sentinel and hard information credit-claiming, she must compare a lottery in which no news, good news, and bad news are all possible (occurring when she deviates from the sentinel's preferred policy) with one in which only no news and good news are possible (when she caters to the sentinel). Which option dominates will depend on the *relative* value of good news and no news, which is encapsulated in the ratio of the voter's posterior beliefs under those two circumstances: $\tilde{\alpha_i}(\text{good news})/\tilde{\alpha_i}(\text{no news})$. This ratio is decreasing in the voter's prior belief about the incumbent, α_i, because the upside potential of good news is lower when the voter already believes that the incumbent is high quality. Hence, an incumbent who is actually low quality when the voters are inclined to believe she is high quality will be most predisposed to "play it safe" by implicitly colluding with the activist group.

3.4 Summary

The implications derived from our simple model concerning the selection and incentive effects of information provision by a sentinel are summarized in Table 3. With respect to the selection mechanism, introducing a sentinel cannot hurt (no revelation), and may possibly help (partial or full revelation), a voter distinguish high- and low-quality candidates for office.[13] Some sentinels will strategically withhold information in pursuit of their own interests, but voters can adjust for this if they understand the sentinel's motives. This serves as a partial confirmation of the conventional understanding of fire alarms as beneficial to a principal in a principal-agent relationship.

Regarding the incentive component of democratic accountability, however, our conclusions are

[13]To be sure, this assumes that the voter processes information rationally. If we relax this assumption — for example by introducing confirmation bias or violations of negative introspection, all bets are off.

Table 3: The Conditional Effects of Third-Party Sentinels on Democratic Accountability

	Cheap Talk		**Hard Information**		**Hard Info + CC**	
Sentinel Type	Revelation	Distortion	Revelation	Distortion	Revelation	Distortion
Neutral	Full	None	Full	None	Full	None
Challenger	None	None	Partial	Sometimes	Full	None
Activist	None	None	Partial	Always	Partial	Sometimes

Notes: (1) Credit-claiming abbreviated as CC; (2) Full revelation means full revelation of available evidence.

less sanguine. Whereas the presence of sentinels won't distort the incentives of the incumbent if sentinel communications are cheap talk, the presence of a sentinel may indeed distort incentives when sentinels can present hard information to voters. Distortions take the form of an incumbent choosing a policy not because it is best ex ante from the voter's perspective, but because it minimizes the probability of bad headlines. In essence, the incumbent may implicitly collude with the sentinel. Allowing the incumbent a channel to publicize her positive accomplishments (using hard information) can eliminate the distortion when the sentinel is a challenger, but not necessarily if he is an interested party.

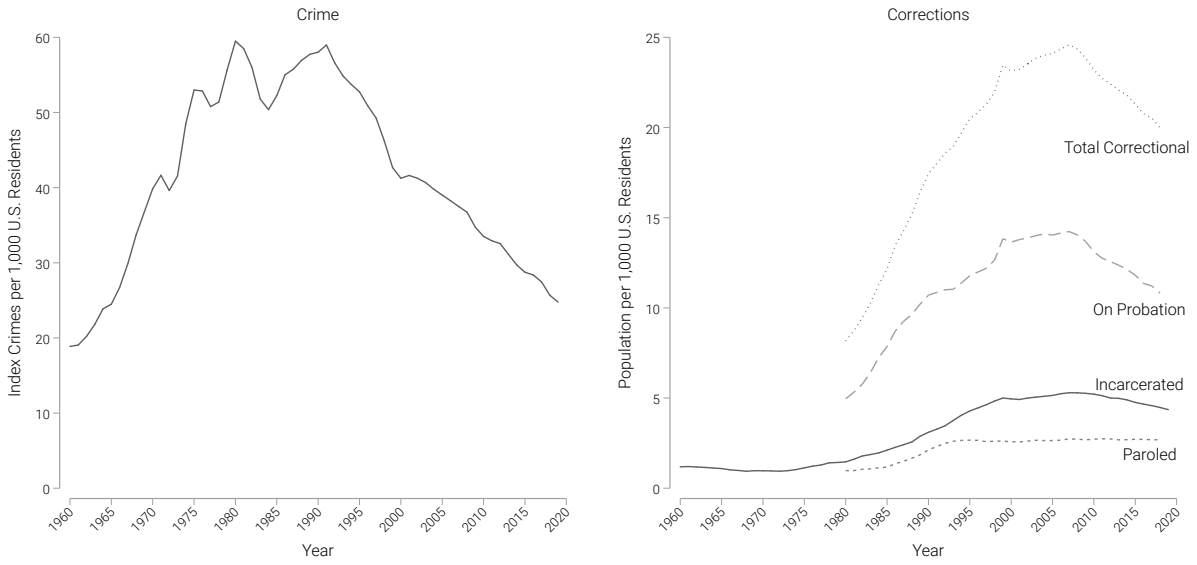
4 Example: The Informational Environment of Criminal Justice Policy

In this section, we argue that a change in the information available to a sentinel – the media – helped drive a change in the political incentives involving criminal justice policy in the early 21st century.[14] While we did not explicitly consider the media in our analysis of sentinels above, they are clearly *the* major conduit through which information is passed. And critically, news outlets are heavily incentivized by the profit motive to prioritize sensational stories with dramatic visuals and emotionally powerful narratives. In the realm of criminal justice policy, this has historically led to an emphasis on bad news, as captured by the cynical cliché, "If it bleeds it leads."

The left panel of Figure 1 displays crime rate data from the FBI Uniform Crime Reporting data base, and corrections data from the Bureau of Justice Statistics. The period between 1960 and 1990 saw a huge increase in crime in the United States, rising from an index crime rate of somewhat less than 20 incidents per 1,000 U.S. residents to nearly 60 from the early-1960s to late 1980s. The policy response is hinted at in the correctional statistics displayed in the right panel: starting in the

[14]The importance of the media in generating political incentives cannot be understated – see, especially, the review essay by Huber and Tucker and the analysis of media congruence and democratic accountability of Canes-Wrone and Kistner, both in this volume.

Figure 1: Crime and Correctional Supervision in the United States, 1960-2019



Crime rate data derived from FBI Uniform Crime Reporting Statistics (`https://www.bjs.gov/ucrdata/abouttheucr.cfm`). Correctional Data (1980-2019) from Bureau of Justice Statistics, Key Statistics (`https://bjs.ojp.gov/data/key-statistics`), with earlier data from "Prisoners 1925-1981", a BJS Bulletin (`https://www.bjs.gov/content/pub/pdf/p2581.pdf`).

mid- to late-1970s, a dramatic increase in the fraction of U.S. citizens behind bars or under some other form of correctional supervision. The burden of incarceration has fallen disproportionately on communities of color.

The 20-year period from 1990 to 2010, however, was characterized by the odd juxtaposition of a marked decline in crime with the enactment, at the local, state, and national levels, of ever-more punitive criminal justice policies. Famously, the 1994 Crime Bill passed with bipartisan support (including the support of a majority of the Congressional Black Caucus, although [Hinton et. al. 2016] notes serious reservations). At the state level, provisions such as truth-in-sentencing laws (requiring convicted felons to serve the full term, or some minimal proportion, of the sentence received at trial), mandatory minimum sentences, the abolition or severe curtailment of parole, and sentencing enhancements for repeat offenders, all contributed to the sharp increase in the proportion of American citizens under some form of correctional control (incarceration, probation, and parole). This proportion, around 8 individuals per 1,000 in 1980, peaked at nearly 25 per in 2007 (more recent data, from 2018 put the figure at around 20 per 1,000).

Coincident with the rise in incarceration was an increased willingness at the local level to employ policing practices such as stop, question, and frisk and militarized police tactics such as SWAT deployment and warrantless, no-knock raids ([Mummolo 2018]). The antecedents of these developments are numerous, including asset forfeiture laws and the reliance of localities on fines and fees as sources of revenue ([Goldstein et. al. 2018]), which create high-powered incentives to police aggressively, particularly in disadvantaged communities of color ([Department of Justice 2015]). Federal law and policies adopted by federal agencies also contributed: the 1994 Crime bill provided funds for hiring 100,000 new police officers. The drug war, combined with the cheap availability of surplus military hardware, contributed to militarization ([Balko 2013]). And the Justice Department's Equitable Sharing program (suspended under the Obama administration but reinstituted under President Trump) allocated a portion of assets seized by federal law enforcement to state and local agencies, even in states that do not have asset forfeiture laws of their own.

Our model illuminates some of the *political* incentives for public officials to support such tough-on-crime policies, and how those incentives may have shifted in the past decade. A key lesson from our analysis is that elected officials may make policy decisions that deviate from the welfare of their citizens in order to minimize the risk that a sentinel will have bad news to pass on to voters. This risk is determined in the model by the sentinel's incentives, and also by the fact that policy choices differ in the probability with which they generate verifiable news (as summarized by the ratio of the parameters π_1 and π_0.) Now suppose, using the language of our model, that $a = 0$ denotes a lenient policy and $a = 1$ denotes a punitive one. Historically, sensational news about crime has meant stories about crime victimization, neighborhood blight, and recidivism, and *not* stories about the social costs of incarceration or aggressive policing. In other words, π_0 was far larger than π_1. In the presence of a sentinel incentivized to publicize bad news, public officials — be they mayors, legislators, or elected prosecutors or judges — all had electoral incentives to push for more punitive policies to avoid the political risk of a harmful headline.

Over the last decade, we have gradually seen an apparent loosening of these incentives: progressive prosecutors have been elected in places like St. Louis, Philadelphia, San Francisco, and Durham. Large majorities of American favor reform of the U.S. criminal justice system. With bipartisan support, Congress passed, and President Trump signed into law, the First Step Act of 2018, which contained a host of sentencing and prison reform provisions. And Civil asset forfeiture

reform has support from both sides of the aisle among lawmakers.

What changed? While a number of factors are clearly at play, one is a mssive increase in the availability of hard information about the downsides of punitive criminal justice policy. Clearly, it has been increasingly difficult to ignore the effects of mass incarceration on communities of color. But technology has also played a significant role of bringing sustained attention to the adverse consequences of incarceration and aggressive policing, beyond those communities. A critical development is the ubiquity of cell phone cameras, which has brought instances of police-initiated violence into the public eye in an unprecedented way. In the language of the model, π_0 may have remained relatively fixed, but π_1 has increased dramatically. Accordingly, officials have reason to fear the political consequences of both kinds of "bad news," which alleviates the political pressure on policymakers toward ever greater levels of punitiveness.

5 Conclusion

As a discipline, political science has long been preoccupied with the implications of an electorate composed in large part of uninformed citizens. In coming to terms with this hard fact of political life, many have pointed to elites as an imperfect solution: by providing cues, elites may help "rationally ignorant" voters vote *as if* they were truly informed. In this paper, we have examined a subtly different role for elites in the political information business. Specifically, we have sought to clarify how third party "sentinels" with the ability to convey information about incumbent performance to voters in the form of "fire alarms" can serve to enhance or undermine democratic accountability. Our analysis reveals three critical lessons.

First, if the information conveyed by a sentinel is unverifiable cheap talk, then only sentinels understood to be neutral, unbiased conduits of information to voters can credibly communicate information about incumbent performance to voters. By contrast, fire alarms sounded by challengers who hope to tarnish the reputation of the incumbent, or interested parties who hope to burnish the reputation of their favored policy irrespective of its utility, will tend to be ignored by voters when those messages are not accompanied by hard evidence.

Second, if a sentinel's message contains hard information, then the presence of a challenger or interested party sentinel may serve to *undermine* democratic accountability, by convincing low quality incumbents to minimize bad news, even if doing so means undermining voter welfare. In the

presence of these biased sentinels, there is an inherent tradeoff between these distortionary incentive effects, which hurt voters, and a richer informational environment, which helps them select good incumbents.

Third, the possibility of incumbent credit claiming can mitigate the distortions induced by biased sentinels, but only under some circumstances. Specifically, distortions will dissipate in the presence of a challenger who only reports bad news about incumbent performance to voters, as they can be countered by the strong incentive of the incumbent to report good news. But distortions may persist when the sentinel is an interested party such as an activist or advocacy group. This is because occasions will emerge in which a low quality incumbent can collude with the sentinel to suppress bad news about her performance. The price is the willingness, on occasion, of low quality incumbents to cater to the sentinel's wishes even when it is not in the voter's interest.

A natural question to ask concerns the possibility of fire alarm accountability in a polarized age. Are voters looking for competent representatives, or simply ideologically well-aligned ones? If the latter, then the limitations of the model we have described above are obvious. And yet, we have no reason to believe that in a different model, in which the uncertainty of the voter concerned whether the incumbent was a moderate or a true believer, all distortions would disappear. As noted above, developing such a model is a task we leave to future research. And yet, at the very least, we can think of our analysis as a cautionary note concerning what kinds of information third parties may bring to bear to help inform voters and create the right incentives for incumbents.

Finally, our model suggests several implications concerning the potential distortions created by biased sentinels. Two that are particularly important concern the availability of challengers and interested parties to serve as sentinels, and the informational environment that governs the production of tangible evidence about policy performance. With respect to the former, we would anticipate fire alarm oversight provided by challenger sentinels to be reduced in safe districts. On the one hand, the standard intuition is that the absence of viable challengers would undermine the accountability of the incumbent. However, this deleterious effect may be mitigated by the removal of the distortionary fear of challengers as self-interested bearers of bad news. Likewise, when thinking about the incentives of incumbents to cater to the demands of interested parties, we must consider what the local interest group and advocacy environment looks like from the perspective of the incumbent: is it, for example, dominated by a single large industry, such that

the incumbent will have a strong incentive to cater to its state-independent policy preferences; or is it rich and heterogeneous, such that it more resembles the case of the neutral conduit. With respect to the latter, we have seen in the case of criminal justice policy how changes in the informational environment may dramatically change the incentives of incumbents.

References

Arnold, R. Douglas. 1990. *The Logic of Congressional Action.* Yale University Press.

Arnold, R. Douglas. 1993. "Can Inattentive Citizens Control their Elected Representatives?" pp. in Bruce Oppenheimer (ed) *Congress Reconsidered.* 5th edition. Washington, DC: CQ Press.

Ashworth, Scott. 2012. "Electoral Accountability: Recent theoretical and Empirical Work." *Annual Review of Political Science* 15 (2012): 183-201.

Ashworth, Scott and Kenneth Shotts. 2011. "Challengers, Democratic Contestation, and Electoral Accountability." APSA 2011 Annual Meeting Paper.

Ashworth, Scott and Kenneth Shotts. 2014. "Challengers and Electoral Accountability," working paper.

Balko, Radley. 2013. *Rise of the Warrior Cop.* Philadelphia, PA: PublicAffairs.

Bawn, Kathleen, et. al. 2021. "Finding a Champion: Principal Agent Relationships in US House Nominations." This Volume.

Beim, Deborah, Alexander V. Hirsch, and Jonathan P. Kastellec. 2014. "Whistleblowing and Compliance in the Judicial Hierarchy." *American Journal of Political Science* 58, no. 4: 904-918.

Besley, Timothy. 2006. *Principled Agents? The Political Economy of Good Government.* Oxford University Press on Demand.

Besley, Timothy, and Andrea Prat. 2006. "Handcuffs for the Grabbing Hand? Media Capture and Government Accountability." *American Economic Review* 96.3: 720-736.

Brookman, David. 2016. "Approaches to Studying Policy Representation," *Legislative Studies Quarterly* 41(1):181-215.

Buisseret, Peter. 2016. "Together or Apart"? On Joint versus Separate Electoral Accountability." *The Journal of Politics* 78.2: 542-556.

Calvert, Randall L. 1985. "The value of biased information: A rational choice model of political advice." *The Journal of Politics* 47(2): 530-555.

Canes-Wrone, Brandice, Michael C. Herron, and Kenneth W. Shotts. 2001. "Leadership and pandering: A theory of executive policymaking." *American Journal of Political Science*: 532-550.

Converse, Philip E. 1964. "The Nature of Belief Systems in Mass Publics." In (eds) Ideology and Discontent pp. 206-261.

Cross, Frank B., and Emerson H. Tiller. 1998. "Judicial partisanship and obedience to legal doctrine: Whistleblowing on the federal courts of appeals." *The Yale Law Journal* 107, no. 7: 2155-2176.

Dellis, Arnaud. 2005. "Blame-Game Politics in a Coalition Government," *Journal of Public Economics* 91:77-96.

Demirkaya, Betul. 2019. "What is Opposition Good For?" *Journal of Theoretical Politics* 31(2):260-280.

Dewan, Torun, and Rafael Hortala-Vallve. 2019. "Electoral Competition, Control and Learning." *British Journal of Political Science* 49, no. 3 (2019): 923-939.

Dewatripont, Mathias, and Jean Tirole. 1999. "Advocates." *Journal of Political Economy* 107, no. 1: 1-39.

Duggan, John, and César Martinelli. 2017. "The political economy of dynamic elections: Accountability, commitment, and responsiveness." *Journal of Economic Literature* 55.3: 916-84.

Dynes, Adam, & John Holbein. 2020. "Noisy Retrospection: The Effect of Party Control on Policy Outcomes," *American Political Science Review*, 114(1), 237-257.

Enelow, James and Melvin Hinich. 1984. *An Introduction to the Spatial Theory of Voting.* Cambridge UP.

Fearon, James D. 1999. "Electoral Accountability and the Control of Politicians: Selecting Good Types versus Sanctioning Poor Performance." Pp. 55-97 in Adam Przeworski, Susan Stokes, and Bernard Manin (eds) *Democracy, Accountability, and Representation,* Cambridge UP.

Fiorina, Morris P. 1981.*Retrospective voting in American National Elections.* Yale University Press.

Fowler, Anthony, Seth Hill, Jeff Lewis, Chris Tausanovitch, Lynn Vavreck, and Christopher Warshaw. 2021. "Moderates". Working paper.

Fowler, Anthony, and Andrew B. Hall. 2018. "Do Shark Attacks Influence Presidential Elections? Reassessing a Prominent Finding on Voter Competence." *The Journal of Politics* 80, no. 4: 1423-1437.

Fox, Justin, and Richard Van Weelden. 2012. "Costly Transparency." *Journal of Public Economics* 96.1-2: 142-150.

Gailmard, Sean. 2014. "Principal Agent Theory and Accountability." The Oxford handbook of public accountability.

Gailmard, Sean, and John W. Patty. 2017. "Participation, Process and Policy: the Informational Value of Politicised Judicial Review." *Journal of Public Policy* 37, no. 3 (2017): 233-260.

Goldstein, Rebecca, Michael W. Sances, and Hye Young You. 2018. "Exploitative Revenues, Law Enforcement, and the Quality of Government Service." *Urban Affairs Review* 56(1): 5-31.

Gordon, Sanford and Gregory Huber. 2002. "Citizen Oversight and the Electoral Incentives of Criminal Prosecutors," *American Journal of Political Science* 46:334-351.

Gordon, Sanford C., and Gregory Huber. 2007. "The Effect of Electoral Competitiveness on Incumbent Behavior." Quarterly Journal of Political Science 2(2): 107-138.

Gordon, Sanford C., Gregory Huber, and Dimitri Landa. 2007. "Challenger Entry and Voter Learning." *American Political Science Review* 101(2):301-320.

Gratton, Gabriele. 2015. "The Sound of Silence: Political Accountability and Libel Law." *European Journal of Political Economy* 37: 266-279.

Hardin, Russell. 2006. "Ignorant Democracy." *Critical Review* 18, no. 1-3: 179-195.

Hinton, Elizabeth, Julilly Kohler-Hausmann, and Vesla Weaver. 2016. "Did Blacks Really Endorse the 1994 Crime Bill?" *New York Times*, April 13.

Hirsch, Alexander V. and Jonathan P. Kastellec. 2019. "A Theory of Policy Sabotage," Working Paper, CalTech Division of Humanities and Social Sciences.

Key, Valdimer Orlando. 1966. *The Responsible Electorate*. Belknap Press of Harvard University Press.

Kishishita, Daiki. 2019. "An Informational Role of Supermajority Rules in Monitoring the Majority Party's Activities." *Journal of Public Economic Theory* 21, no. 1: 167-196.

Lemon, Andrew Y. 2005. *Reputational Concerns in Political Agency Models.* Doctoral disseration, Yale Department of Economics (available on line at http://www.princeton.edu/~smorris/past%20PhD%20Students)

Maskin, Eric, and Jean Tirole. 2004. "The Politician and the Judge: Accountability in government." *American Economic Review* 94.4 (2004): 1034-1054.

Mayhew, David. 1974. *Congress: The Electoral Connection*. Yale University Press.

McCubbins, Mathew D., and Thomas Schwartz. 1984. "Congressional Oversight Overlooked: Police Patrols Versus Fire Alarms." *American Journal of Political Science*: 165-179.

Milgrom, Paul R. 1981. "Good News and Bad News: Representation Theorems and Applications." *Bell Journal of Economics* 12(2): 380-391.

Milgrom, Paul. 2008. "What the Seller Won't Tell You: Persuasion and Disclosure in Markets." *Journal of Economic Perspectives* 22, no. 2: 115-131.

Morelli, Massimo and Richard Van Weelden. 2013. "Ideology and Information in Policy Making," *Journal of Theoretical Politics* 25(2):412-39.

Mummolo, Jonathan. 2018. "Militarization fails to enhance police safety or reduce crime but may harm police reputation." *Proceedings of the National Academy of Sciences* 115(37): 9181-9186.

Pitkin, Hanna F. 1967. *The Concept of Representation*. University of California Press.

Prat, Andrea. 2005. "The Wrong Kind of Transparency." *American Economic Review* 95.3: 862-877.

Prendergast, Canice. 2003. "The Limits of Bureaucratic Efficiency." *Journal of Political Economy* 111, no. 5: 929-958.

Schultz, Kenneth A. 1998. "Domestic Opposition and Signaling in International Crises." *American Political Science Review* 92.4: 829-844.

Snyder, James M. Jr., and Michael M. Ting. 2003. "Roll Calls, Party Labels, and Elections," *Political Analysis* 11(4): 419-444.

Society for Civil Engineers. 2017. "2017 Infrastructure Report Card: Infrastructure Scores a D+," https://www.infrastructurereportcard.org/

Stasavage, David. 2007. "Polarization and Publicity: Rethinking the Benefits of Deliberative Democracy." *The Journal of Politics* 69.1: 59-72.

Warren, Patrick L. 2012. "Independent Auditors, Bias, and Political Agency." *Journal of Public Economics* 96.1-2: 78-88.

Wilson, J.Q., 1989. *Bureaucracy: What Government Agencies Do and Why They Do It.* Basic Books.

Wolton, Stephane. 2019. "Are Biased Media Bad for Democracy?" *American Journal of Political Science* 63, no. 3: 548-562.

United States Department of Justice, Civil Rights Division. 2015. "Investigation of the Ferguson Police Department." Available at `http://www.justice.gov/sites/default/files/opa/press-releases/attachments/2015/03/04/ferguson_police_department_report.pdf`.