2023

A Theory of Claim Resolution

Scott Baker
Washington University in St. Louis School of Law

Lewis A. Kornhauser
New York University School of Law

A Theory of Claim Resolution

Scott Baker*
Washington University in St. Louis, School of Law, One Brookings Drive, St. Louis, MO 63130, USA

Lewis A. Kornhauser
New York University, School of Law, 40 Washington Square South, New York, NY 10012, USA

We study claim resolution. A claim consists of a global fact and a local fact. The global fact is observed by the principal and the agent. The local fact is observed by the agent alone. The agent resolves the claim; the principal decides whether the agent is more likely wrong or right. The principal and agent can disagree about the weight to accord each fact or the overall evidence threshold. The agent cares whether the principal follows or ignores her advice. We characterize how the equilibrium varies with the nature of disagreement. Despite lacking commitment power, we find that the principal grants the agent decision-making authority over an interval of global facts. Further, we find that the principal can better motivate an agent who excessively weights the local fact than an agent who excessively weights the global fact. The principal strictly prefers the former to the latter even though either would make the same number of errors if granted complete autonomy. (JEL C7, K0, D7, K4, M4)

1. Introduction

Judges, legal scholars, and philosophers often disagree about the “method” by which court should decide cases. In contract, for example, there is a long-standing debate between “formalists” and “anti-formalists” (Corbin 1964–1965; Charny 1999; Scott 1999; Bernstein 2015). Formalists believe that the text of the contract should take primacy in establishing the rights and obligations of the parties. The court should thus rarely consider extrinsic evidence of the parties’ intent, such as trade usage or oral testimony. Formalists, in short, are wedded to the plain meaning rule, which states:

> [W]hen the provisions in the contract are clear and unambiguous, the court looks only to the four corners’ of the document

*We thank the editor, three referees, Andy Daughety, John Ferejohn, and Claudio Mezzetti for helpful comments on this project. Participants at workshops at the University of Virginia, New York University, George Mason University, the American Law and Economics Conference, and the Law and Economic Theory Workshop held at Duke University provided much fruitful feedback.

The Journal of Law, Economics, and Organization, Vol. 39, No. 1
doi:10.1093/jleo/ewab017
Advance Access Publication December 10, 2021

> in arriving at the intent of the parties. In the absence of any ambiguity, the contract will be enforced according to its terms because no construction is appropriate.[1]

Anti-formalists, in contrast, believe the court should dive into the overall context of the agreement, including the manner under which the parties performed under the contract (as evidence of what they believed the contract means) and how members of the trade conduct business. Unlike formalists, anti-formalists place much less weight on the text of the contract itself. Indeed, one of the most famous anti-formalist scholars writes:

> [N]o man can determine the meaning of written words by merely glueing his eyes within the four corners of a square paper; to convince that it is men who give meanings to words and that words in themselves have no meaning; and to demonstrate that when a judge refuses to consider relevant extrinsic evidence on the ground that the meaning of written words is to him plain and clear, his decision is formed by and wholly based upon the completely extrinsic evidence of his own personal education and experience (Corbin 1964–1965: 164).

This same debate extends beyond contracts to statutory and constitutional interpretation. And it reverberates within the law itself. For example, in the United States, the statute governing the sale of goods has much more of an anti-formalist bent than the common law of contracts (Goetz and Scott 1985: 274).

Disagreement over the method is not limited to the courts. In the loan context, the literature has debated the agency costs associated with loan officer discretion and whether loans should be made on hard information alone or some combination of hard and soft information (Liberti and Mian 2009; Godbillon-Camus and Godlewski 2013; Liberti and Petersen 2019). In that context, we see the same debate about the method. Loan officers and supervisors might disagree about how much weight should be accorded to the soft information about the applicant such as personal connections and trust and how much to hard information, such as financial wherewithal.

This article considers the implication of disagreement over method for judges, loan officers, and other actors who disagree about the weight to be accorded different pieces of evidence. We ask what happens when: (a) actors in a hierarchy disagree about the method as well as outcomes; (b) the reviewing actor cannot commit to a policy of reversal, and (c) the agent pays a cost when her advice is ignored (i.e., she is reversed).

In the model, the agent must make a dichotomous decision subject to the principal's subsequent review or oversight. The agent has access to two pieces of information. Specifically, the agent observes a global fact

[1] *Amoco Prod. Co. v. EM Nominee P'ship Co.*, 2 P.3d 534, 540 (Wyo. 2000).

and a local fact which bear on the decision. In contrast, the principal only observes the global fact. In a contract dispute, for example, the trial judge sees the demeanor of the witness testifying to the trade usage and the contract text, whereas the appellate court only sees the text itself. The text of the contract could be ambiguous, for example, while the witness's demeanor (and testimony) clearly points to liability.

In such a setting, the agent and the principal might disagree about method—the weight to be allocated each piece of evidence. They might also disagree about outcomes—the sum total of evidence necessary to declare the claim "valid."

The literature on delegation and cheap talk focuses on this second source of disagreement (Crawford and Sobel 1982; Holmstrom 1984; Alonso and Matouschek 2008). The first source of disagreement is, we believe, more novel.

We first investigate what decisions the principal will reverse and why. Specifically, how does the reversal decision depend on interaction between the location of the global fact and the source of agency conflict? Does it matter whether the conflict arises from differences of opinion over method or differences of opinion over outcomes? And if so, why?

In the equilibrium where the agent's decision conveys information about the local fact, the principal affirms unless the agent's decision is both unexpected given the information contained in the global fact and *sufficiently* more in line with the agent's than the principal's preferences. It is not enough, in other words, for the principal to know that the agent improperly places a thumb on the scale in favor of valid decisions. The global fact must also be informative enough on its own (e.g., it must unambiguously point to invalidity) to allow the principal to make a credible threat to reverse. That means the principal—who has no commitment power—delegates broad swathes of decisions to the agent who holds different views. In other words, the agent gets his preferred outcome even though the principal has not given the agent any real authority.

In fact, the principal often affirms the agent's decision even when it is contrary to what the principal would have decided if forced to do so on her own. The agent has information the principal lacks. By granting discretion to the agent to decide as she sees fit, the principal willingly pays the price of the agent deciding some cases in a way the principal disfavors to leverage the agent's information for other cases where the agent and principal share the same goal. In this latter set of cases, the principal's decision would misfire if based on the global fact alone; that is without the benefit of the agent's knowledge of the local fact.

We further show that the equilibrium when the principal faces an agent who overweights the local fact (relative to what the principal prefers) differs dramatically from the equilibrium when the agent underweights the local fact.

Following the law literature, we characterize an agent who overweights the local fact as anti-formalist. The anti-formalist, for example, places too little weight on the contractual text and too much weight on other harder-to-observe markers of contractual intent. In contrast, the agent who under-weights the local fact is a formalist.

An anti-formalist agent will, on occasion, decide a case as valid and other times invalid. She will also trigger scrutiny and face reversal threats for both types of decisions. Formalist agents will also, on occasion, decide a case as valid and other times decide a case as invalid. Contrasted with anti-formalists, the formalist agent often faces no reversal threat whatsoever. Further, in circumstances where they do face a reversal threat, it is only with respect to one type of decision, The principal will (a) always affirm the formalist agent's valid decisions, or (b) always affirm the agent's invalid decisions, or (c) affirm every decision the formalist agent makes.

After characterizing the principal's review strategy and the agent's resolution strategy, the welfare implications of the model are discussed. Like Gennaioli and Shleifer (2007), our model sits in a two-dimensional space, consisting of a local fact and a global fact. Preferences are defined by cutlines in this space. The principal and the agent's cutlines have a slope and an intercept. As such, the preference conflict between the principal and the agent can, as noted, take multiple forms. And this innovation reveals a new tradeoff in the selection of agents. The anti-formalist agent—because she faces a credible threat of reversal in more cases—is easier to motivate than agents exhibiting any other type of preference conflict. Because of this, the principal strictly prefers the anti-formalist agent, even if that agent's preferences are less congruent with his own than other agents he might select. This result stands in contrast to the ally principle touted in political science that states "[i]f the boss delegates, then she picks the agent whose ideal point is the closest to hers" (Bendor et al., 2001: 243).

In short, not all differences of opinion between the agent and the principal matter in the same way. The principal can effectively manage some preference conflicts—specifically regarding an agent's tendency to overweight local facts—better than others.

The welfare implications of the model speak directly to the type of frontline loan officer a bank should hire. The bank should prefer an agent who cares more deeply about soft information than the principal does even if that agent disagrees more with the supervisor than other potential hires.

What about judges? Of course, unlike a supervisor of loan officers, the appellate court does not pick the trial court judges in the federal system. Nonetheless, the model surfaces new tradeoffs in the appointment process. Take a president looking to appoint a new judge. Assume he cannot get his first best choice through the Senate confirmation process. The model shows that a judge's philosophy matters in different ways depending on the place the judge sits in the judicial hierarchy. Anti-formalist trial court judges, in general, have less power over the ultimate resolution of cases. Formalist trial judges, in contrast, provide little useful information when they resolve cases and rarely face reversal threats. The president, thus, might pick an anti-formalist nominee to the trial court level with whom he substantially disagrees over a formalist nominee whose preferences are more congruent with his own. That same tradeoff does not appear for Supreme Court justices or circuit court appointees.

Finally, our model bears on the value of commitment in matters of claim resolution. In claim resolution, the principal seeks to minimize findings of validity where she prefers invalidity and findings of invalidity where the principal prefers validity. The number of correct answers is irrelevant.

This preference feature dampens the value of commitment. Indeed, we find that the principal has the same payoff whether she can commit to delegate to the agent certain classes of decisions or must engage in *ex post* reversal after observing the agent's decision. While the agent reacts differently in these two settings, those differences do not change the overall number of mistakes made in equilibrium. Instead, it shifts the composition as between the types of mistakes.

The article unfolds as follows. A review of the related literature follows immediately. Section 2 sets up the model. Section 3 characterizes the equilibrium of the model. Section 4 provides the welfare and value of commitment results. Section 5 discusses in detail the relationship between our model and other canonical models in the literature. Section 6 provides a short conclusion. All proofs not in the text can be found in the Appendix.

1.1 Related Literature

First, the model builds off past work examining the interactions between appellate courts and trial courts. Cameron et al. (2000), for example, look at the Supreme Court's decision whether to grant certiorari and thereafter overrule an appellate court decision. In that model, the Supreme Court and lower court can only disagree along one-dimension: the threshold of proof. The model thus cannot explore implications of method disagreement among judges, which is our focus.[2]

We share a two-dimensional case space model with other prior work in the courts literature. That said, we ask substantially different questions. For instance, Gennaioli and Shleifer (2007) examine the welfare gains when a judge "distinguishes" a prior precedent. The core insight is that distinguishing improves the efficiency of the judge-made law. In contrast, we ask about the ability of trial courts to communicate information to appellate courts through their resolutions. Lax (2012) asks whether the appellate court should craft a rule or a standard when facing a potentially hostile lower court. Our appellate court cannot commit to an *ex ante* legal rule or standard. Instead, it must react to the decision of the trial court. Again, the issue is about communication from a trial court to the appellate court. Finally, Bueno de Mesquita and Stephenson (2002) focus on communication from the appellate court to the trial court. The issue they

[2] Likewise, disagreement along a single dimension defines the tax compliance literature (Andreoni et al., 1998). In these models, the taxpayer always prefers to report less rather than more income, an assumption we relax in our article.

study is this: when should an appellate court judge break with precedent rather than refine it? Breaking with precedent leads to a less informed decision by the trial court (she ignores all the prior precedent signals set by the past appellate court judges), but can lead to a decision closer to the writing judge's ideal point. In that model, the trial court is assumed to be a faithful agent; as such it neglects the central tension in our model.

Second, our work uses as a scaffold the classic model of Crawford and Sobel (1982). We pivot from the core assumptions of that model in a few ways. One, as this is a model of claim resolution, the agent's message space and principal's action space are restricted. The agent can only send a valid or invalid message. The principal is restricted to accept or reverse the agent's recommendation. In equilibrium, for any given global fact, the agent slices the space of local facts into two sets: local facts that signal validity and local facts that signal invalidity. Furthermore, the relative size of these sets differs in intuitive ways with the location of the global fact. For example, if the global fact suggests invalidity is the proper action, the agent must partition to ensure that the set of local facts pointing to validity is smaller than the set of local facts pointing to invalidity.

Finally, we find that the agent's partition leads the principal to adopt the agent's preferred outcome when the global fact is sufficiently uninformative or the conflict between the principal and agent is sufficiently small. In the cheap talk model, in contrast, the principal rarely takes the exact action that the agent prefers. In this way, our model extends the results of the delegation literature where the agent obtains his desired outcome over some interval (Holmstrom 1984; Manuel and Bagwell 2013) to a class of problems where the principal lacks commitment power.

Third, we extend the literature that identifies the benefits and costs of preference conflict between a principal and an agent. For example, Che and Kartik (2009) explore what happens when the agent carries different priors from the principal. The authors show that this agent will often work harder to find information as a result, and, therefore, can be of greater value to the principal. This same theme arises in Aghion and Tirole (1997) as to the allocation of formal versus real authority in organization, Dewatripont and Tirole (1999) as to the benefits of advocacy, and Gennaioli and Shleifer (2007) on the value of polarization in a judiciary.

Our agent does not exert effort. Instead, the principal's reversal threat motivates certain kinds of agents more than others. The threat is more effective with an agent who holds a specific type of preference conflict: when the agent is both antiformalist (weights the local fact too much) and, on average across all claims, equally likely to make a mistaken finding of validity or a mistaken finding of invalidity.

2. The Model

The model involves a principal and an agent. In our motivating courts example, the principal is the appellate court and the trial court is the agent.

In the banking example, the principal is the supervisor and the agent is the frontline loan officer.

At the start of the game, the agent is presented with a "claim." A claim consists of two facts: a global fact x and a local fact y. The global fact is observable to both the principal and the agent. The local fact is private information, observable only by the agent.

The global fact and the local fact are randomly drawn from independent, uniform distributions with support on $[0, 1]$. The space of the possible claims is thus the unit square.

When presented with a claim, the agent decides whether to find the claim valid ("1") or invalid ("0"). The agent's strategy is a function $d = \Delta(x, y)$ specifying for each possible claim whether she will decide the claim as valid.

The principal observes the agent's decision and the global fact. Based on these two pieces of information, the principal must decide whether to reverse ("1"), affirm ("0"), or mix between the two actions. The principal's strategy is thus a function $\gamma = g(d, x)$ that specifies the probability of reversal for each possible agent decision and location of the global fact.

Together the decisions by the principal and agent yield a final resolution of the claim $r = \rho(d, \gamma)$, where

$$\rho(d, \gamma) = \gamma(1 - d) + (1 - \gamma)d.$$

The principal and the agent care about the final resolution. In addition, as noted in Section 1, the agent cares about reversal—an aspect of her utility we discuss in a moment.

Cutlines partition the space of claims into ones that the agent or principal prefers to find valid and ones that they prefer to find invalid.

The principal's partition is

$$\frac{x}{2} + \frac{y}{2} = \frac{1}{2}.$$

The principal equally weights the local and global facts in any decision. Moreover, the weighted sum of the evidence must exceed 1/2 for the principal to prefer validity. Rearranging, the principal divides the claim space with a cutline of $y = 1 - x$. She prefers validity if $y \geq 1 - x$ and invalidity if $y < 1 - x$. In contrast, the agent's partition is

$$wx + (1 - w)y = z$$

where $0 \leq w < 1$ is the weight accorded the global fact and $0 < z < 1$ is the total amount of evidence the agent needs to find validity. With these parameters, the agent's cutline is

$$f(x) = \frac{z}{1 - w} - \frac{wx}{1 - w}.$$

The agent prefers validity if $y \geq f(x)$ and invalidity if $y < f(x)$.

The agent and principal might disagree about method—the weight w to accord local versus global facts—or the threshold of evidence z needed to impose liability: that is, the burden of proof.

On the one hand, the agent might be more of an anti-formalist than the principal ($w < 1/2$). Such an agent cares too deeply about the local fact, like oral testimony by the parties. On the other hand, the agent might be more of a formalist than the principal ($w > 1/2$). This agent, for example, places too much import on the text of the contract than the principal would prefer. And, of course, anti-formalist and formalist agents can disagree more or less as to the burden of proof to deploy: that is, they can be lax or strict.

Figure 1 depicts the cutlines for the principal and the agent where the agent is an anti-formalist ($w < 1/2$) and strict ($z > 1/2$). The blue line is the principal's cutline, the green line is the agent's cutline, and $x_c = (1 - w - z)/(1 - 2w)$ marks the global fact where the principal and agent's cutlines cross. In the areas marked as I or II, the principal and agent agree on the disposition (invalid in areas marked as I; valid in areas marked as II). Areas III and IV measure the degree of disagreement between the principal and the agent. In Area III, the agent prefers the case be found valid where the principal prefers invalidity. In Area IV, the principal prefers validity and the agent prefers invalidity.

The sum of Areas III and IV indexes the amount of disagreement between the principal and the agent. Area III can be computed by subtracting the small right triangle from the larger right triangle; that is,

$$\begin{aligned}\text{Area(III)} &= \frac{x_c^2}{2} - \frac{x_c}{2}\left(\frac{z}{1-w} - (1 - x_c)\right)\\ &= \frac{x_c}{2}\left(x_c - \frac{z}{1-w} + 1 - x_c\right)\\ &= \frac{x_c}{2}\left(\frac{1-w-z}{1-w}\right)\\ &= \frac{(1-w-z)^2}{2(1-2w)(1-w)},\end{aligned}$$

using that $x_c = (1 - w - z)/(1 - 2w)$. Likewise, Area IV can be computed by subtracting the smaller right triangle from the larger one.

$$\begin{aligned}\text{Area(IV)} &= \frac{(1-x_c)^2}{2} - \frac{1-x_c}{2}\left(1 - x_c - \frac{z-w}{1-w}\right)\\ &= \left(\frac{1-x_c}{2}\right)\left(1 - x_c - (1 - x_c) + \frac{z-w}{1-w}\right)\\ &= \frac{(z-w)^2}{2(1-2w)(1-w)},\end{aligned}$$

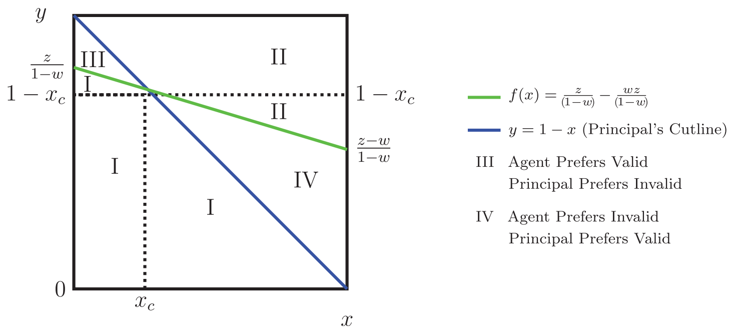

y
$\frac{z}{1-w}$
$1-x_c$
III
I
II
II
I
IV
I
$1-x_c$
$\frac{z-w}{1-w}$
0
x_c
x
$f(x) = \frac{z}{(1-w)} - \frac{wz}{(1-w)}$
$y = 1 - x$ (Principal's Cutline)
III Agent Prefers Valid Principal Prefers Invalid
IV Agent Prefers Invalid Principal Prefers Valid

Figure 1. Principal-Agent Preference Conflict.

using the fact that $1 - x_c = (z - w)/(1 - 2w)$. Eventually, in Section 4, the analysis will turn to the principal choosing among similarly situated agents where "similar" agents present the same amount of disagreement. That choice reveals insights about the relative costs and benefits of different ways an agent might disagree with the principal.

Each player suffers a loss of 1 from an error in claim resolution. The error might be a mistaken finding of validity or a mistaken finding of invalidity. Because they have different cutlines, the players disagree about what counts as an error.

The principal's payoff is

$$U_p(x, y, r) = -rI_{y<1-x} - (1 - r)(1 - I_{y<1-x}),$$

where I is an indicator function that takes on a value of 1when $y < 1 - x$ Suppose that $y < 1 - x$ and, thus the principal prefers the claim be decided as invalid. If the claim is resolved as valid ($r = 1$) the principal suffers from a mistaken resolution. On the other hand, if the claim is decided as invalid, no mistake is made and the principal suffers no loss.

The agent's payoff is the sum of her loss from mistakes in the final resolution and the loss in the event she is reversed. The reversal loss is

$$c(y, d) = \begin{cases} ky & \text{if } d = 0 \\ k \times (1 - y) & \text{if } d = 1 \end{cases}$$

where k is a constant. We offer the following interpretation to justify this loss from reversal.

(1) Reduced Form for Reputational Harm. Reversal might hurt the agent because other actors—future employers or, perhaps with our courts example, the US Supreme Court—see the reversal and think the agent is incompetent. As the clarity of the evidence for, say, a valid resolution increases, we suspect that the agent will have an easier time convincing a third party that she was correct

> and the principal was incorrect. The result blunts the stigma of reversal. This might happen if the agent could produce a signal for third parties correlated with the local fact to explain why the principal was wrong to reverse.[3]

Given this reversal cost, the agent's payoff is

$$U_a(x, y, r, d, \gamma) = -rI_{y<f(x)} - (1-r)(1 - I_{y<f(x)}) - \gamma c(y, d).$$

The parameters of the payoff functions, the location of the global fact, and the distributions from which the claims arise are common knowledge. The only thing the agent knows that the principal does not is the location of the local fact.

The framework is expansive enough to allow for multiple kinds of agency conflict. The conflict can arise out of differences of opinion as to method, the threshold of proof, or both. Table 1 presents the possible agency conflicts:

Let $b(y|d, x)$ be the principal's posterior belief about the location of y given the agent's decision d and the global fact x. A perfect Bayesian equilibrium consists of a vector of strategies, $(\Delta^\star, g^\star)$, and posterior beliefs, $b^\star$, such that:

(1) For each claim, given beliefs $b^\star(\cdot)$ the principal's reversal policy, $g^\star(\cdot)$, solves

$$\max_{\gamma} \int_0^1 U_p(x, y, \rho(d, \gamma)) b^\star(y|d, x) dy$$

(2) For each claim, given $g^\star$ the agent's decision policy $\Delta^\star(\cdot)$ solves:

$$\max_{d} U_a(x, y, \rho(d, \gamma^\star), d, g^\star(x, d))$$

(3) On the equilibrium path, and, to the extent possible, off the equilibrium path, the principal's beliefs, $b^\star$, are formed according to Bayes' Rule.

In a perfect Bayesian equilibrium, the agent resolves the claim optimally given the equilibrium reversal strategy of the principal. The principal reverses when doing so maximizes her expected utility given her posterior beliefs about y. Finally, the principal's posterior beliefs are derived from the agent's equilibrium strategy using Bayes' rule.

Following real-world claim resolution practice (e.g., appellate review of trial courts), we view this game as one where the agent resolves the claim and the principal must decide whether to reverse or affirm.

[3] For notational simplicity, we assume that the loss from reversal is independent of x. The appendix shows that the equilibrium analysis remains the same if (a) the agent and principal's losses from mistaken adjudication linearly increases with the size of the error and (b) the agent suffers a fixed cost of reversal.

Table 1. Possible Sources of Agency Conflict

	Lax	Strict	Agree on burden
Anti-formalist	$w < \frac{1}{2}, z < \frac{1}{2}$	$w < \frac{1}{2}, z > \frac{1}{2}$	$w < \frac{1}{2}, z = \frac{1}{2}$
Formalist	$w > \frac{1}{2}, z < \frac{1}{2}$	$w > \frac{1}{2}, z > \frac{1}{2}$	$w > \frac{1}{2}, z = \frac{1}{2}$
Agree on method	$w = \frac{1}{2}, z < \frac{1}{2}$	$w = \frac{1}{2}, z > \frac{1}{2}$	$w = \frac{1}{2}, z = \frac{1}{2}$ (No conflict)

Alternatively, the game can be seen as a cheap talk game where the agent sends one of two messages (valid or invalid). That is to say, the agent makes a recommendation to the principal as to how the claim should be resolved. The principal can follow the recommendation or not. The agent's recommendation is costless (all that matters for payoffs is the ultimate resolution imposed by the principal), but the agent suffers a reputational loss when her advice is ignored.

3. Equilibrium

We focus on the equilibrium where the agent's decision conveys information to the principal about the location of the local fact.[4] The principal and agent's preferences are partially aligned; that is, there will be a set of cases where the principal and agent agree on the outcome. The trouble is that the agent's resolution of the claim sends a noisy signal about the location of the local fact. It might be a claim where the principal and agent agree or it might be a claim where the principal and agent disagree. The size of these two sets depends on the location of the global fact and the agent's equilibrium strategy.

In doing the analysis, it is fruitful to separate formalist agents ($w > 1/2$) from anti-formalist agents and those agents that agree with principal on method ($w \leq 1/2$). Why? A geometric insight provides the answer.

For any global fact, the agent's cutline might lie below, above, or equal to the principal's cutline. As Figure 2 illustrates, for the anti-formalist agent, the cutline will lie below the principal's to the left of where the principal and agent's cutlines cross. For a formalist agent, the opposite is true. And this difference matters for the equilibrium.

In the figure, the black area represents cases where the agent prefers valid and the principal prefers invalid, whereas the orange area represents cases where the agent prefers invalid and the principal prefers valid.

Imagine that the global fact is 0. The principal observes this global fact. She also knows the preferences of the agent. Notably, for a case with a global fact of 0, the anti-formalist might draw a local fact where she

[4] The Appendix discusses equilibria where the agent's decision is unrelated to the location of the local fact and thus uninformative. If they exist, these equilibria fail the "universal divinity" refinement of Banks and Sobel (1987).

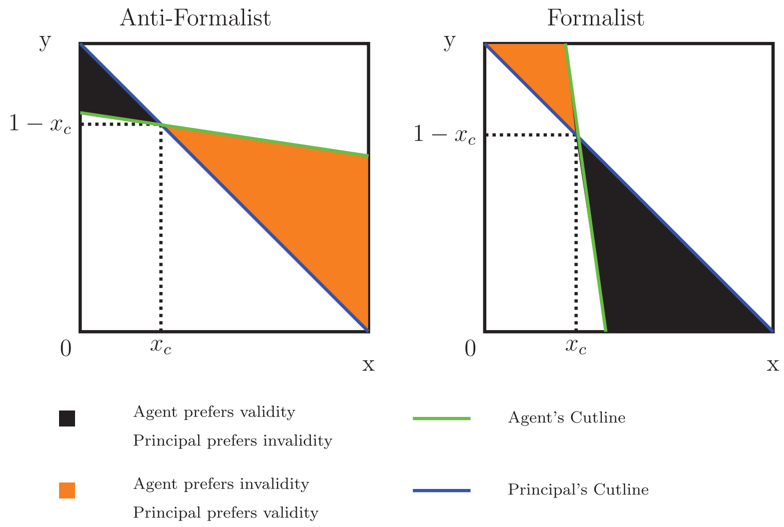


Figure 2. Preference Conflicts.

prefers validity and the principal does not (a case in the black area at $x = 0$). In contrast, at $x = 0$, the formalist agent will never draw a local fact where she prefers validity and the principal does not. Instead, the only potential conflict arises if the formalist agent draws a local fact where she prefers invalidity and the principal prefers validity (a case in the orange area).

If $x = 0$, the principal can credibly threaten to reverse valid decisions by the agent. After all, the principal knows that her own preference is for invalidity when $x = 0$. In contrast, the principal cannot credibly threaten to reverse invalid decisions. The reason is that the global fact favors invalidity.

As a result of the principal's inability to make a credible threat with respect to invalid resolutions, the formalist agent gets her preferred resolution when the global fact is located at 0. In contrast, the principal can lodge a credible threat of reversal against the anti-formalist agent for this same global fact. This difference in the credibility of threats of reversal implies that formalist and anti-formalists exhibit divergent kinds of behavior in equilibrium, which we explore more fully below.

3.1 Anti-Formalist Agents and Agents Who Agree on Method

Consider first what happens when an anti-formalist agent wants to find more claims valid than the principal. Specifically, the global fact is less than x_c and thus the potential conflict involves the black area in Figure 2.

Assume the agent decides according to her cutline: she finds the claim valid ($d = 1$) when $y \geq f(x)$ and invalid otherwise. Following the valid resolution of case involving global fact x, the principal believes the local fact is distributed uniformly with a support $[\max\{0, f(x)\}, 1]$. The max function takes account that the agent's cutline might lie below 0. If it does, the valid decision by the agent deciding according to her cutline provides no information to the principal. The principal continues to believe the local fact is distributed uniformly on [0, 1].[5]

If the principal affirms the agent's valid decision, the final resolution is valid ($r = 1$). Now we know that

$$U_p(x, y, 1) = \begin{cases} -1 & \text{if } y < 1 - x \\ 0 & \text{otherwise.} \end{cases}$$

Accounting for the principal's updated beliefs about y, her expected payoff from affirming the agent's valid decision is

$$\begin{aligned} \frac{\int_{\max\{0,f(x)\}}^{1} U_p(x, y, 1)\mathrm{d}y}{\mathrm{pr(valid)}} &= -\frac{\int_{\max\{0,f(x)\}}^{1-x} \mathrm{d}y + \int_{1-x}^{1} 0\mathrm{d}y}{\mathrm{pr(valid)}} \\ &= -\frac{1 - x - \max\{0,f(x)\}}{\mathrm{pr(valid)}}. \end{aligned} \tag{1}$$

If instead, the principal reverses the agent's valid decision, the final resolution is invalid ($r = 0$). We know that

$$U_p(x, y, 0) = \begin{cases} 0 & \text{if } y < 1 - x \\ -1 & \text{otherwise.} \end{cases}$$

Therefore, the principal's expected payoff from reversing is

$$\begin{aligned} \frac{\int_{\max\{0,f(x)\}}^{1} U_p(x, y, 0)\mathrm{d}y}{\mathrm{pr(valid)}} &= -\frac{\int_{\max\{0,f(x)\}}^{1-x} 0\mathrm{d}y + \int_{1-x}^{1} \mathrm{d}y}{\mathrm{pr(valid)}} \\ &= -\frac{x}{\mathrm{pr(valid)}}. \end{aligned} \tag{2}$$

The principal affirms a valid decision if Equation (1) exceeds Equation (2). If $x = 0$, the principal's loss from reversing a valid decision is 0, and thus she would certainly do so. With that in mind, we can locate the smallest global fact where the principal prefers to affirm a valid decision by the agent, while still recognizing that any valid final resolution comports perfectly with what the agent desires. That value is determined by

$$\frac{\int_{\max\{0,f(x)\}}^{1} U_p(x, y, 1)\mathrm{d}y}{\mathrm{pr(valid)}} - \frac{\int_{\max\{0,f(x)\}}^{1} U_p(x, y, 0)\mathrm{d}y}{\mathrm{pr(valid)}} = 0.$$

[5] As an example, take an agent whose cutline is $f(x) = -x$. This agent prefers to find all claims valid. If the principal observes a valid resolution from this agent, the principal's beliefs do not change as to the local facts that might give rise to that decision. It could be any local fact.

or,

$$-\frac{1-x-\max\{0,f(x)\}}{\mathrm{pr(valid)}}+\frac{x}{\mathrm{pr(valid)}}=0. \tag{3}$$

After plugging in $f(x)=z/(1-w)-wx/(1-w)$ solve Equation (3) for x. The solution marks a lower bound on the global fact,

$$\underline{x}=\min\left\{\frac{1}{2},\frac{1-w-z}{2-3w}\right\}. \tag{4}$$

If $x\geq\underline{x}$, the global fact does not contain enough evidence of invalidity for the principal to credibly threaten reversal of a valid decision. The agent takes advantage of this informational deficiency to disregard the principal's preferences entirely.

What happens when $x<\underline{x}$? If the agent decided every case as she preferred, the principal would reverse valid resolutions. But then the agent would want to deviate to avoid the reversal cost. In this circumstance, the equilibrium strategy for the principal involves mixing between affirming and reversing a valid decision by the agent. On the other hand, the agent's strategy is a cutoff point $y^\star$.

Specifically, the agent finds the claim valid if $y\geq y^\star$ and invalid otherwise. The point $y^\star$ lies in the interval $(f(x),1-x)$. Intuitively, the agent moderates her behavior to find more claims invalid than she wants to. But she does not perfectly follow the principal's cutline.

The principal understands the agent's cutoff point. Indeed, that point partitions the space of local facts as between valid and invalid resolutions in a particular fashion. Given the agent's cutoff point, it must be that the principal is equally likely to make a mistake when she affirms and when she reverses a valid resolution. If that is true, the principal is willing to mix. The probability of reversal, then, induces the agent to select that point.

After observing a valid resolution, the principal believes that the local fact is uniformly distributed on $[y^\star,1]$. Given these beliefs, the difference in the principal's expected payoff from affirming and reversing is

$$\frac{\int_{y^\star}^{1}U_p(x,y,1)\mathrm{d}y}{\mathrm{pr(valid)}}-\frac{\int_{y^\star}^{1}U_p(x,y,0)\mathrm{d}y}{\mathrm{pr(valid)}}=-\frac{\int_{y^\star}^{1-x}\mathrm{d}y}{\mathrm{pr(valid)}}+\frac{\int_{1-x}^{1}\mathrm{d}y}{\mathrm{pr(valid)}}$$

To induce mixing, the equilibrium strategy of the agent, $y^\star$ must make the principal indifferent. Or

$$-\frac{1-x-y^\star}{\mathrm{pr(valid)}}+\frac{x}{\mathrm{pr(valid)}}=0 \tag{5}$$

Observe the solution to Equation (5) is $y^\star(x)=1-2x$, which we now write as a function of x to make plain that the agent's equilibrium cutoff point changes with the global fact.

Next, the agent must be willing to play this cutoff point, given the reversal probability γ. As noted, the only agents who might decide the case as valid are those who draw local facts above their cutline, that is in the interval $[f(x), 1]$. Such an agent has two choices. First, they might decide the case as invalid and be affirmed for sure. If they do so, they suffer a loss of 1. Alternatively, the agent might decide the claim as valid and hope to be affirmed. This course of action results in a loss $\gamma \times (1 + k(1 - y))$. Set these two values equal to locate the indifferent agent type.

$$-1 + \gamma \times [1 + k(1 - y)] = 0 \tag{6}$$

Plug in $y^\star(x) = 1 - 2x$ into Equation (6) and solve for γ. The solution identifies the principal's mixing strategy, $\gamma^\star = g^\star(x, 1) = \frac{1}{1+2xk}$.

Finally, the agent suffers a lower cost of reversal as she becomes more confident in the correctness of her decision. That means the agent prefers validity for local facts above $y^\star$ and invalidity for local facts below $y^\star$.

To sum up, Equation (3) defines a marker between global facts with complete deference to a valid decision by the agent and cases where the principal can make a credible reversal threat. The joint solution to Equations (5) and (6) identifies the equilibrium behavior $(y^\star, \gamma^\star)$ for cases with global facts less than $\underline{x}$.

Now take a global fact where the anti-formalist agent prefers to find more claims invalid than the principal, the orange area in Figure 2. Again, there will be a range of global facts where the information content of the global fact is too weak for the principal to effectively threaten reversal even if she knows the agent is acting solely in her own private interest.

Suppose, as before, the agent decides all cases as she prefers. Following an invalid decision by the agent, the principal believes the local fact is distributed on the interval $[0, \min\{f(x), 1\}]$. The principal's payoff to affirming the invalid decision and having a final resolution of invalid ($r = 0$) is

$$\begin{aligned}\frac{\int_0^{\min\{f(x),1\}} U_p(x, y, 0)\mathrm{d}y}{\mathrm{pr(invalid)}} &= -\frac{\int_{1-x}^{\min\{1,f(x)\}} \mathrm{d}y}{\mathrm{pr(invalid)}} \\ &= -\frac{\min\{1, f(x)\} - (1 - x)}{\mathrm{pr(invalid)}}.\end{aligned} \tag{7}$$

If the principal reverses an invalid decision by the agent, the final resolution is valid ($r = 1$). The resulting expected payoff to the principal is

$$\begin{aligned}\frac{\int_0^{\min\{f(x),1\}} U_p(x, y, 1)\mathrm{d}y}{\mathrm{pr(invalid)}} &= -\frac{\int_0^{1-x} \mathrm{d}y}{\mathrm{pr(invalid)}} \\ &= -\frac{1 - x}{\mathrm{pr(invalid)}}.\end{aligned} \tag{8}$$

The principal will affirm the agent's invalid decision if Equation (7) exceeds Equation (8). We can thus define the largest value of x where the principal will affirm an invalid decision by an agent who decides all cases according to her own cutline. That value occurs when

$$-\frac{\min\{1, f(x)\} - (1-x)}{\text{pr(invalid)}} + \frac{1-x}{\text{pr(invalid)}} = 0, \tag{9}$$

from which we solve for an upper bound.

$$\overline{x} = \max\left\{\frac{1}{2}, \frac{2-z-2w}{2-3w}\right\} \tag{10}$$

If $x \leq \overline{x}$, the principal affirms any invalid resolution. If $x > \overline{x}$, the equilibrium involves mixing by the principal and a cutoff point by the agent.

For these cases, suppose the agent plays a cutoff point, $y^\star$, which is larger than $1-x$. The difference in the principal's payoff from affirming and reversing an invalid agent decision is

$$\frac{\int_0^{y^\star} U_p(x,y,0)\mathrm{d}y}{\text{pr(invalid)}} - \frac{\int_0^{y^\star} U_p(x,y,1)\mathrm{d}y}{\text{pr(invalid)}} = -\frac{\int_{1-x}^{y^\star} \mathrm{d}y}{\text{pr(invalid)}} + \frac{\int_0^{1-x} \mathrm{d}y}{\text{pr(invalid)}}$$

For the principal to mix, she must be indifferent given the agent's cutoff point strategy, $y^\star$. Meaning

$$-\frac{y^\star - 1 - x}{\text{pr(invalid)}} + \frac{1-x}{\text{pr(invalid)}} = 0. \tag{11}$$

Likewise, the agent must be willing to play the cutoff point $y^\star$ given the reversal probability, γ. The agent finds it optimal to play this strategy when

$$-1 + \gamma \times [1 + ky^\star] = 0 \tag{12}$$

For claims with global facts above $\overline{x}$, the joint solution $(y^\star, \gamma^\star)$ to Equations (11) and (12) identifies the equilibrium. The first proposition summarizes formally the discussion thus far.

Proposition 1. If $w \leq 1/2$, there exists an equilibrium consisting of the triple $(\Delta^\star, g^\star, b^\star)$ such that:

1. $$\Delta^\star(x,y) = \begin{cases} 1 & \text{if } y > y^\star(x) \\ 0 & \text{if } y \leq y^\star(x). \end{cases}$$

where

$$y^{\star}(x) = \begin{cases} 1-2x & \text{if } x \in [0, \underline{x}) \\ f(x) & \text{if } x \in [\underline{x}, \overline{x}] \\ 2(1-x) & \text{if } x \in (\overline{x}, 1]. \end{cases}$$

2. $$\gamma^{\star} = g^{\star}(x,d) = \begin{cases} \dfrac{1}{1+2xk} & \text{if } x < \underline{x} \text{ and } d = 1 \\ \dfrac{1}{1+2k-2kx} & \text{if } x > \overline{x} \text{ and } d = 0 \\ 0 & \text{otherwise.} \end{cases}$$

3. Beliefs about y are uniform with support $[0, \min\{y^{\star}(x), 1\}]$ if the decision is invalid and support $[\max\{y^{\star}(x), 0\}, 1]$ if the decision is valid.

Proof. The proof follows by solving Equations (3) and (9) for $\underline{x}$ and $\overline{x}$. The joint solution to Equations (5) and (6) define the equilibrium for cases with global facts less than $\underline{x}$. Next, the joint solution to Equations (11) and (12) defines the equilibrium for global facts greater than $\overline{x}$. Finally, given $y^{\star}(x)$, it is clear that the principal prefers to affirm all invalid resolutions below $\overline{x}$ and all valid resolutions above $\underline{x}$. □

Two examples illustrate the insights from Proposition 1.

Example 1 (Pure Anti-Formalist). Suppose that w = 0 and $z = 1/2$. This agent agrees with the principal that half of the claims should be held valid and half invalid. But the agent disagrees as to the method. She thinks the global fact should be ignored. As a result, *ex ante* the agent and principal disagree about the resolution in 1/4 of the claims.

Plugging this agent's parameter values into Equations (4) and (10) provides the markers on the interval of discretion. Specifically, the upper and lower bounds on the global facts are

$$\underline{x} = \frac{1}{4}$$
$$\overline{x} = \frac{3}{4}.$$

Outside these bounds, the principal mixes when the agent makes an unexpected decision (i.e., a decision that goes against what the global fact suggests is the right decision). The agent's cutoff point is set to provoke the principal's indifference.

Figure 3 illustrates the preference conflict and equilibrium strategies for the agent and the principal. To draw this figure and all the remaining ones, we set $k = 1$.

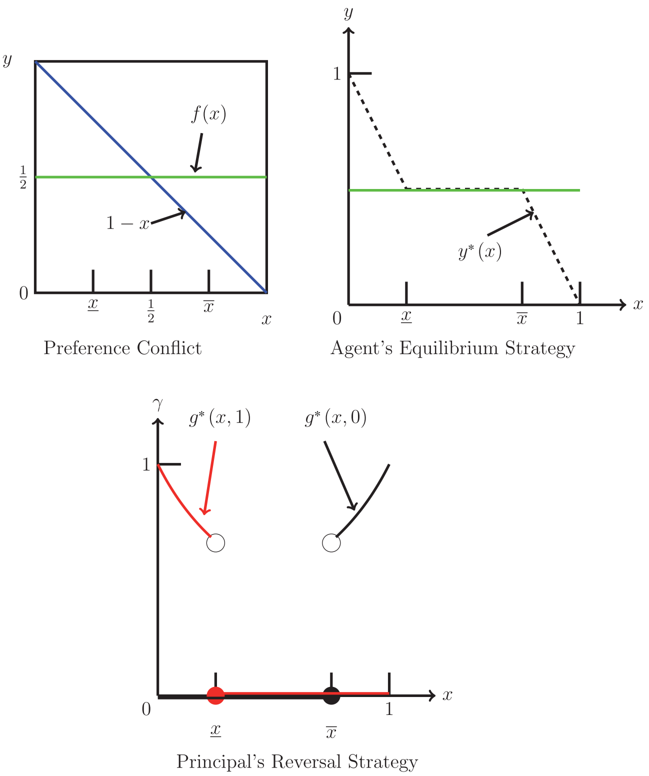
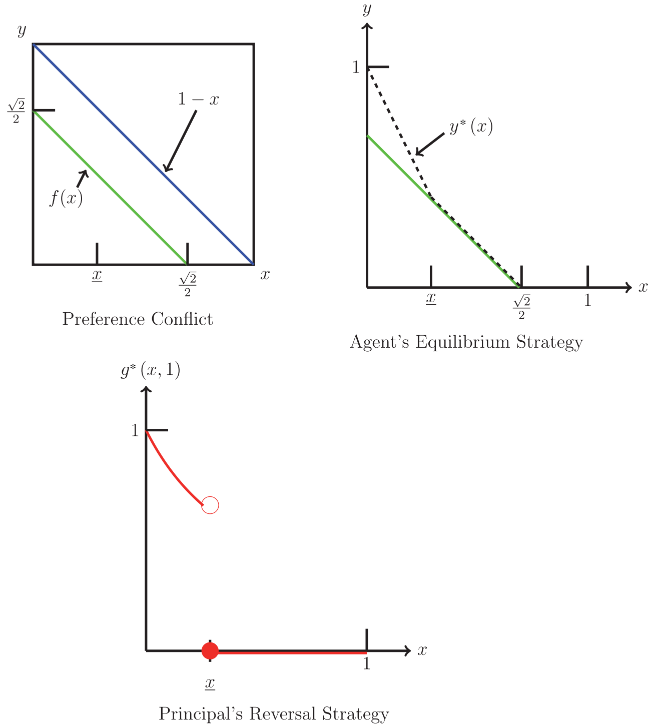


Figure 3. Pure Anti-Formalist.

Between the markers $\underline{x}$ and $\overline{x}$, the principal affirms all decisions. Below $\underline{x}$, the principal affirms invalid decisions and reverses valid decisions with positive probability. The principal does so because the global fact is small and, therefore, counter to the agent's resolution. As the global fact approaches 0, the principal reverses valid decisions with greater frequency.

Likewise, for cases with global facts above $\overline{x}$, the principal affirms all valid resolutions and reverses invalid resolutions with positive probability. Notably, the jump in the reversal probability at $\underline{x}$ and $\overline{x}$ arises because of the nature of the payoffs. The agent's loss from a mistaken resolution is a fixed cost of 1. To induce an agent who realizes a local fact above his cut-line to choose invalidity and suffer a loss for sure rather select validity and suffer a loss with some probability demands a large reversal probability punch.

Example 2 (Agent Disagrees About Burden Alone). Suppose that $w = 1/2$ and $z = \sqrt{2}/4$. The preference conflict and equilibrium strategies appear in Figure 4. Like the pure anti-formalist, in this setting, the agent and principal disagree about the resolution in 1/4 of all cases.



Preference Conflict

Agent's Equilibrium Strategy

Principal's Reversal Strategy

Figure 4. Threshold Disagreement.

Use these parameters and Equations (4) and (10) to obtain the bounds on the global facts

$$\underline{x} = \frac{2 - \sqrt{2}}{2} \approx .29$$

$$\overline{x} = \frac{4 - \sqrt{2}}{2} \approx 1.29$$

Because the upper bound exceeds 1, the principal defers to all invalid decisions made by this agent, meaning $g(x, 0) = 0$ for all x. This makes sense. Whenever the agent prefers an invalid resolution the principal does too. As a result, the agent only faces a reversal threat with respect to valid decisions. And in fact below $\underline{x}$, the agent and principal's behavior mirror the equilibrium behavior of Example 1.

Proposition 1 offers four lessons. First, in many cases, the principal defers to every decision the agent makes. And this happens despite the principal's inability to commit to do so and the fact that the global fact points against the agent's decision. Meaning, the principal would have decided differently if she had to make the call on her own.

The model thus sheds light on the instructions among judges at different levels of a hierarchy. According to the US Supreme Court,

> If the district court's account of the evidence is plausible in light of the record viewed in its entirety, the court of appeals may not reverse it even though convinced that had it been sitting as the trier of fact, it would have weighed the evidence differently. *Anderson v. City of Bessemer City*, N.C., 470 U.S. 564, 573–74 (1985).

Appellate courts are not to reverse, even if they would have decided the case differently. How is that commitment possible? This longstanding practice arises in our model: the global fact is inconsistent with the district court decision, and yet the appellate court affirms. The key is that the appellate court trades off the agent's informational advantage against the potential preference conflict. For certain cases, the appellate court infers that the trial court was more likely right than wrong in spite of observing that the trial court ruled contrary to the weight of the global fact: the text of the contract, for instance.

Second, appellate courts are often instructed to affirm unless the trial court decision is clearly erroneous. In this model, we see an arguable—and indeed subtle—difference between clear errors and run-of-the-mill errors by trial courts. A naive way to characterize error is when a trial court decision goes against the weight of the publicly available information. Our model demonstrates that the appellate court understands that this does not necessarily mean the trial court's made a mistake. Indeed in many cases, the appellate court believes the trial court's decision is more likely right than wrong, even though the appellate court would have decided differently if forced to do so on the global fact. Outside the bounds of discretion, the trial court moderates its behavior so that the principal believes the trial court's decision equally likely to be right or wrong.

Yet, in reversing the appellate court has access to a piece of information to justify its decision. The appellate court can report in the opinion that the trial court was clearly erroneous. Why? The appellate court can point to the fact that the global fact presented strong evidence of, say, validity while the trial court found the claim invalid. Interestingly, the appellate court can make that statement, knowing that the trial court's decision is, in equilibrium, equally likely to be right or wrong.

Third, the agents described in the two examples disagree *ex ante* with the principal in the same percentage of the cases. Yet the principal can lodge a credible reversal threat in more cases against the anti-formalist agent. This suggests, and we will confirm in the welfare section, that the principal strictly prefers the anti-formalist agent even when this agent's preferences are less aligned with the principal's than other potential agents.

Fourth, the model predicts that some agents will face two bounds on permissible behavior, while others will face only one. Return here to our motivating example of the frontline loan officer. Suppose she weighs local facts more heavily than her superior, but is neither biased in favor nor against granting a loan. Such an agent will face supervisor scrutiny as to the improper grant of loans and the improper denial of loans. That is to say, the denial of a loan to an applicant who has a strong credit score will trigger supervisor review and potential reversal. Likewise, the grant of a loan to an applicant with a weak credit scores will trigger review and potential reversal. In contrast, if the loan officer simply has a tendency to grant too many applications, the supervisor will only scrutinize the grant of loans to applicants with weak credit scores.

3.2 The Formalist Agent: $w > 1/2$

Having considering the anti-formalist and agents that agree on method, we next turn to the formalist agent. As noted, the formalist agent's cutline will lie below the principal's to the right of where they cross and above to the left. First focus on what happens when the agent's cutline lies below the principal's; that is, the $f(x) < 1 - x$ the agent prefers to find more claims valid than the principal.

Suppose the agent does not moderate her behavior. The principal's payoff to affirming a valid decision is Equation (1). The payoff to reversing is Equation (2). The payoff to affirming exceeds the payoff to reversing in two cases:

$$x > \frac{1}{2} \text{ or } f(x) \in [1 - 2x, 1 - x]$$

As shown in Figure 2, from the principal's perspective, the formalist agent finds too many claims valid as the global fact increases. If $x > 1/2$, the principal lacks the evidence to reverse such a finding and thus affirms. If $x < 1/2$, a valid finding becomes suspect. But there is a competing consideration: the agent's preferences might be in tune with the principal's. Indeed the agent and principal might actually be in agreement about the resolution of cases for that global fact. If so, the principal will want to affirm any finding by the agent. In other words, for global facts near where the principal and agent's preferences are in harmony, the principal always affirms, irrespective of what the global fact suggests is the correct answer. The second condition captures this idea.

Next if a global facts lies below 1/2 and $f(x) < 1 - 2x$, then the equilibrium is identified as the solution $(y^\star, \gamma^\star)$ to Equations (5) and (6). In this range, the global fact suggests invalid is the correct answer and the principal and agent's preferences are in sufficient disharmony to trigger the equilibrium where the principal mixes.

What happens if the agent prefers to find more claims invalid than the principal? Assuming the agent follows his cutline, the principal's payoff to affirming exceeds the payoff to reversing in two cases.

$$x < \frac{1}{2} \text{ or } f(x) \in [1 - x, 2(1 - x)].$$

The logic mirrors the prior discussion. The principal will affirm a invalid decision if

(1) The global fact provides an insufficient basis to overrule the invalid decision or
(2) The principal and agent's preferences are in sufficiently aligned.

To complete this discussion, consider what happens if the global fact lies above 1/2 and $f(x) > 2(1-x)$. For these claims, the equilibrium is defined by the joint solution to and (11).

The next proposition summarizes these points.

Proposition 2. If $w > 1/2$, there exists an equilibrium consisting of the following triple $(\Delta^\star, g^\star, b^\star)$ such that:

-
$$\Delta^\star(x,y) = \begin{cases} 1 & \text{if } y > y^\star(x) \\ 0 & \text{if } y \leq y^\star(x). \end{cases}$$

where

$$y^\star(x) = \begin{cases} 1-2x & \text{if } f(x) < 1-2x \text{ and } x < \dfrac{1}{2} \\ 2(1-x) & \text{if } f(x) > 2(1-x) \text{ and } x > \dfrac{1}{2} \\ f(x) & \text{otherwise.} \end{cases}$$

-
$$\gamma^\star = g^\star(x,d) = \begin{cases} \dfrac{1}{1+2xk} & \text{if } f(x) < 1-2x,\ x < \dfrac{1}{2} \text{ and } d=1 \\ \dfrac{1}{1+2k-2kx} & \text{if } f(x) > 2(1-x),\ x > \dfrac{1}{2}, \text{ and } d=0 \\ 0 & \text{otherwise.} \end{cases}$$

- Beliefs about y are uniform with support $[0, \min\{y^\star(x), 1]$ if the decision is invalid and support $[\max\{y^\star(x), 0\}, 1]$ if the decision is valid.

Proof. Proof follows from discussion in text. □

To explain this proposition, two examples will be helpful.

Example 3 (The Formalist Agent Who Agrees About z). Let $w = 9/10$ and $z = 1/2$. Figure 5 shows the preference conflict and the agent's equilibrium strategy. Notice that $f(x)$ resides between $1-2x$ and $2(1-x)$. As a result, no cases exists, where $f(x) < 1-2x$ and $x < 1/2$. Similarly, no

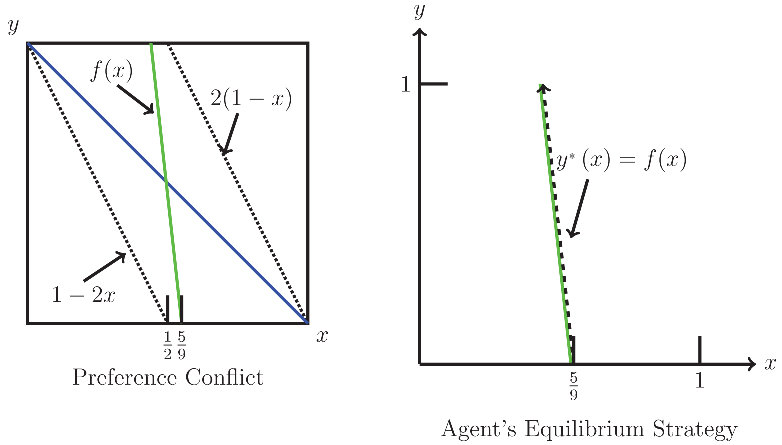


Figure 5. Agent where $w = \frac{9}{10}$ and $z = \frac{1}{2}$.

case exists where $f(x) > 2(1 - x)$ and $x > 1/2$. As a result, this agent always gets her preferred outcome.

Example 4 (The Formalist Agent Who Prefers a Lower Threshold). Let $w = 9/10$ and $z = 1/4$. Notice that $f(x) < 2(1 - x)$ when $x < 1/2$. As a result, this agent moderates her behavior. She faces a reversal threat for global facts between $[\underline{x}, \frac{1}{2}]$ and moderates her behavior accordingly. The preference conflict and equilibrium strategies appear in Figure 6.

The examples show the difference between the formalist and anti-formalist agents. It is impossible for the formalist agent to face reversal threats with respect to more than one type of decision. There is basic geometry behind this statement. To face reversal threats for both valid and invalid decisions the agent's cutline must reside below $1 - 2x$ for a case with global facts less than 1/2 and above $2(1 - x)$ for case with a global fact greater than 1/2. No agent with a cutline whose slope is steeper than the principal's can meet both these conditions. In contrast, the anti-formalist agent often does meet both these conditions.

Taken together, Propositions 1 and 2 segment the parameter space into six buckets as illustrated in Table 2. For each bucket, we indicate whether the equilibrium can involve no bounds (i.e., complete discretion), a lower bound, an upper bound, or both.

This section closes by highlighting differences between this model and the classic signaling models in the literature. The model shares some features with Crawford and Sobel (1982) and the costly signaling models such as Spence (1973). Indeed, it combines elements of both models while not perfectly tracking either.

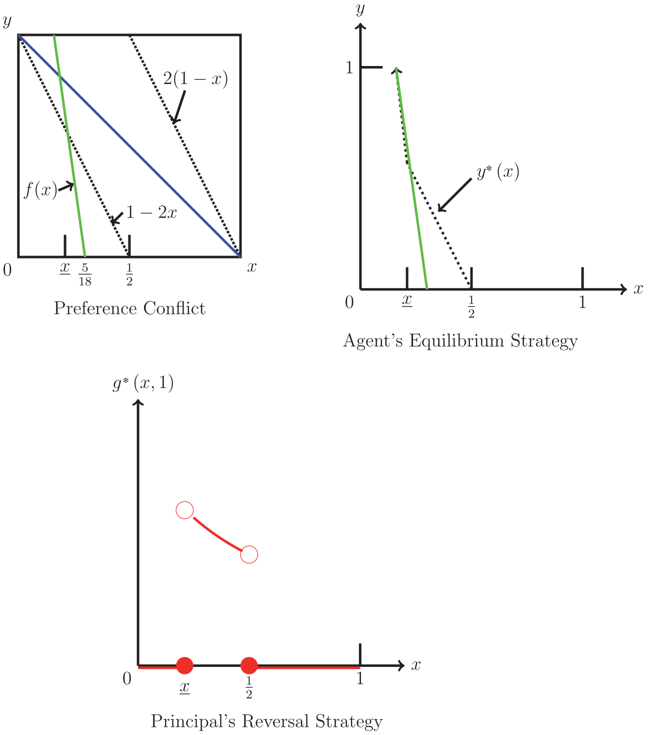


Figure 6. Agent where $w = \frac{9}{10}$ and $z = \frac{1}{4}$.

First, in the cheap talk model, the message is costless to send. In our model, the agent pays a penalty when her advice is ignored. The message is no longer costless. It depends on what the principal does. Meanwhile, in the costly signaling model, the agent always pays the signaling cost irrespective of the beliefs or actions of the principal. In our model, there is no cost to an agent sending a signal if the principal follows the advice.

Second, in our model, the agent and principal do not disagree in every state of the world. No matter the case draw, there will be realizations of the local fact where the principal and agent agree on the outcome. This stands in contrast to the cheap talk model.[6]

At least in the context of claim resolution, our assumptions on bias are more plausible than the standard assumption of bias in every state of the

[6] To see that consider the leading example from Crawford and Sobel (1982). The preferences are given by

$$U_p = -(a-y)^2$$
$$U_a = -(a-y-b)^2$$

where y is the state of the world known only by the agent, a is the action the principal takes and b is the agent's bias. For all y, the agent and principal disagree about the action.

Table 2. Equilibrium Behavior

	Lax $z < 1/2$	Strict $z > 1/2$	Agree $z = 1/2$
Anti-Formalist $w < 1/2$	Lower bound; upper bound; or both	Lower bound; upper bound; or both	Both upper and lower bound
Formalist $w > 1/2$	At most one bound below 1/2	At most one bound above 1/2	No constraint on behavior
Agree $w = 1/2$	Lower bound	Upper bound	No conflict

world. In part, the plausibility of our assumption results from the dichotomous nature of claim resolution. It is implausible to believe that principal and agent disagree on the resolution of every claim; it may be plausible in the standard model which assumes a continuous action and state space that the agent always wants more or less of some action. But claim resolution, as done by courts, loan officers, administrators of social security and veteran's affairs, and parole boards, does not seem to fall into this setting.[7]

Finally, and most importantly, once the principal is convinced that the local fact lies above or below some threshold, more fine-grained information about "how" large or small y actually isdoes not change the principal's decision. In the cheap talk model, the principal always wants to take a higher action when the state is larger. More fine-grained information, then, induces the principal to make different choices. This feature is noticeably absent from our model.

The last difference means that, while we might allow the agent to send more than two messages and then construct equilibria with more fine-grained partitions of local facts, those equilibria do not improve the principal's welfare. The next proposition sheds additional light on this point.

Proposition 3. Assume the anti-formalist agent can send three distinct messages. Equilibria with three distinct messages exist. In any of these equilibria, the principal obtains the same expected welfare as in the two message equilibrium derived in Proposition 1.

4. Welfare and the Value of Commitment

The ally principle from political science (Epstein and O'Halloran 1999; Bendor et al. 2001) suggests that the principal should seek out agents whose preference are most akin to their own. This section reveals that that result does not extend once disagreement can occur along multiple dimensions.

[7] Recall that, in our setting, the agent (and the principal) are resolving claims, not announcing policy. So while it might be, for instance, that an agent always prefers a more claimant-favorable policy than the principal, they may still agree about the resolution of specific claims.

To get the intuition, contrast examples 1 and 3. In Example 1, the anti-formalist agent faces significant reversal threats. Knowing this, she is motivated to partially decide as the principal prefers, making fewer mistakes from the principal's perspective in equilibrium. Example 3 is a formalist who faces no reversal threat whatsoever. Although the formalist agent in Example 3 presents a lower amount of *ex ante* preference conflict with the principal, she prefers the anti-formalist agent because that agent is easier to control and motivate.

Before proceeding to generalize this welfare result, an assumption about the severity of the underlying preference conflict helpfully restricts the parameter values under consideration.

Assumption 1(Limited Disagreement).

- Denote the percentage of claims over which agent and principal have an *ex ante* disagreement as A. Assume that A is less than 1/4.
- The principal and agent's preferences are such that the preference is perfectly aligned for some case within the unit interval. Formally, $x_c \in (0, 1)$.

Without an agent, the principal would decide all cases with global facts below 1/2 as invalid and all cases with global facts above 1/2 as valid, resulting in an error rate of 1/4. The principal can do better than this by employing an agent. But which type? Of course, agents might disagree more or less with the principal. As noted above, we say that two agents are *ex ante* "identical" if they would make the same number of errors in a world where the principal lacked the power to overrule.

Under this definition of similar agents, the principal's program is to select an agent—a pair (w, z)—to maximize her welfare subject to a fixed amount of *ex ante* disagreement.

Facing an anti-formalist, the amount of disagreement is the sum of the Areas III and IV in Figure 1. That sum is

$$\frac{(1-w-z)^2+(z-w)^2}{2(1-w)(1-2w)}.$$

The principal's *ex post* welfare accounts for the amount of compliance by the anti-formalist agent. For global facts in the interval $[\underline{x}, \overline{x}]$, the agent does not moderate her behavior. Thus, the principal's welfare reflects the area of disagreement in that range.

For cases below $\underline{x}$, the agent adopts the cutoff point $y(x) = 1 - 2x$. That choice reduces the area of disagreement. For any x in this range and $y \geq 1 - 2x$, the agent decides the case as valid. If the principal reverses, she suffers a loss if $y \in [1 - x, 1]$, or a loss with probability x. If the principal affirms, she suffers a loss if $y \in [1 - 2x, 1 - x]$, as such she suffers a loss with probability x. As a result, no matter whether she reverses or not, the principal suffers a loss of x.

Had the principal granted the agent complete autonomy, she would have suffered a loss of $(1-x)-f(x)$ for cases in this interval. Given the mitigation, she now suffers a loss of x. And so, for each global fact, the principal's welfare can be computed as the amount of *ex ante* disagreement $(1-x)-f(x)$ less the benefits of mitigation $(1-2x)-f(x)$.

Proceeding this way, over the interval $[0,\underline{x}]$, the principal's welfare is Area III less the benefits of mitigation: the black area in Figure 7.

We get

$$\begin{aligned}\text{Area}(\text{III})-\text{Area}(\text{Black}) &= \frac{(1-w-z)^2}{2(1-w)(1-2w)}-\frac{\underline{x}}{2}(1-f(\underline{x}))-\left(\frac{z}{1-w}-f(\underline{x})\right)\Bigg)\\ &= \frac{(1-w-z)^2}{2(1-w)(1-2w)}-\frac{(1-w-z)^2}{2(1-w)(2-3w)}\\ &= \frac{(1-w-z)^2(2-3w)}{2(1-w)(1-2w)(2-3w)}-\frac{(1-w-z)^2(1-2w)}{2(1-w)(2-3w)(1-2w)}\\ &= \frac{(1-w-z)^2}{2(1-2w)(2-3w)}.\end{aligned} \tag{13}$$

where we used that $\underline{x}=\frac{1-w-z}{2-3w}$.

For cases above $\overline{x}$, we can do the same calculation. Subtract from Area IV, the amount of mitigation, the red area in the figure, revealing:

$$\begin{aligned}\text{Area}(\text{IV})-\text{Area}(\text{Red}) &= \frac{(z-w)^2}{2(1-2w)(1-w)}-\left(\frac{1-\overline{x}}{2}\right)\left(f(\overline{x})-\left(f(\overline{x})-\frac{z-w}{1-w}\right)\right)\\ &= \frac{(z-w)^2}{2(1-2w)(1-w)}-\left(\frac{1-\overline{x}}{2}\right)\left(\frac{z-w}{1-w}\right)\\ &= \frac{(z-w)^2}{2(1-2w)(1-w)}-\frac{(z-w)^2}{2(2-3w)(1-w)}\\ &= \frac{(z-w)^2(2-3w)}{2(1-2w)(1-w)(2-3w)}-\frac{(z-w)^2(1-2w)}{2(2-3w)(1-w)(1-2w)}\\ &= \frac{(z-w)^2}{2(2-3w)(1-2w)}.\end{aligned} \tag{14}$$

Putting Equations (13) and (14) together, the principal's welfare from selecting an antiformalist agent is

$$W=-\frac{(1-w-z)^2+(z-w)^2}{2(2-3w)(1-2w)}. \tag{15}$$

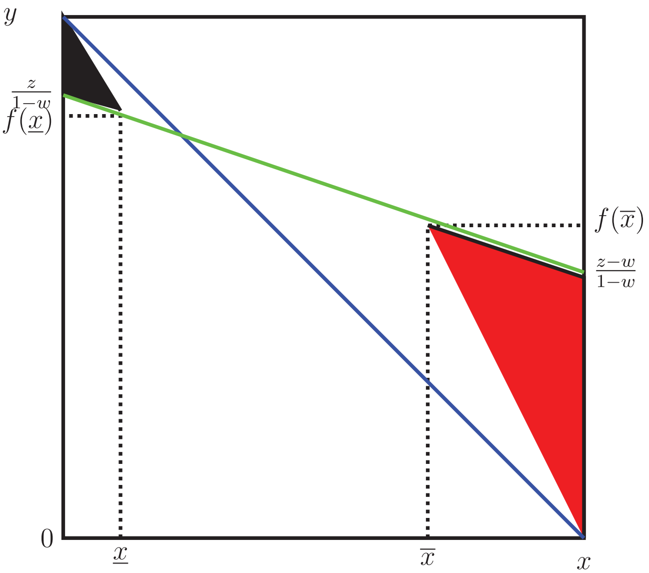


Figure 7. Equilibrium Mitigation.

The principal's program maximizes Equation (15) subject to the constraint

$$\frac{(1-w-z)^2+(z-w)^2}{2(1-w)(1-2w)}=A.$$

In the Appendix, we derive the solution to this program. Given Assumption 1, the solution is $(w,z)=\left(\frac{1-4A}{2-4A},\frac{1}{2}\right)$; this pair characterizes the second-best optimum among anti-formalist agents.

Using the definition of x_c, this second-best optimal agent's cutline crosses the principal's at

$$\begin{aligned}x_c&=\frac{1-w-z}{1-2w}\\&=\frac{\frac{1}{2}-\frac{1-4A}{2-4A}}{1-2\left(\frac{1-4A}{2-4A}\right)}\\&=\left(\frac{2-4A-2(1-4A)}{2(2-4A)}\right)\left(\frac{2-4A}{2-4A-2(1-4A)}\right)\\&=\frac{1}{2}.\end{aligned}$$

In words, the optimal anti-formalist agent is equally likely to commit a mistaken finding of validity or a mistaken finding of invalidity.[8] Inserting this agent's cutline into the definition of welfare, we obtain

$$W = -\frac{A}{1+4A},$$

The question is whether the principal can do better than this by selecting a formalist agent or an agent who disagrees solely as to the evidence threshold. Our next proposition shows that she cannot, and provides the central welfare result of the model.

Proposition 4.

(1) Given Assumption 1, the principal strictly prefers to appoint the anti-formalist agent over an agent who disagrees in the same percentage of cases, but whose disagreement manifests as a dispute about the evidence threshold only.

(2) Given Assumption 1, the principal strictly prefers to appoint the optimal anti-formalist agent over a formalist agent who disagrees in the same percentage of cases.

Proof. See Appendix □

The principal trades enhanced *ex ante* agency conflict for more effective control *ex post*. As the anti-formalist agent becomes more balanced—equally likely to make either type of error—she becomes easier to motivate. The motivation effect mitigates the agency cost from disagreement, making the anti-formalist agent especially attractive to the principal.

An example amplifies the point. Consider the anti-formalist agent from Example 1, defined by the $(w, z) = (0, \frac{1}{2})$ or a cutline $y = 1/2$. *Ex ante*, this agent disagrees in 1/4 of all cases. But she can be motivated through the threat of reversal. As such, by selecting this agent, the principal enjoys welfare of

$$\begin{aligned} W &= -\frac{\frac{1}{4}}{1+1} \\ &= -\frac{1}{8}. \end{aligned}$$

Next, consider a formalist agent defined by $(w, z) = \left(\frac{3}{4}, \frac{1}{2}\right)$. *Ex ante*, this agent disagrees with the principal in 1/6 of the cases. Yet, Proposition 2 teaches that this formalist agent faces no threat of reversal. Thus, the principal obtains welfare of -1/6 by hiring her. Even though the formalist agent is less disagreeable at the outset, the principal strictly prefers to appoint the anti-formalist agent.

[8] The result of preferring an agent who commits equal errors, we suspect, is a product of the principal's cutline being a 45 degree line.

Of course, some formalist agents will partially comply with the principal's wishes, as Example 4 demonstrates. Inspection of Figure 8 provides the logic behind the proof of the desirability of anti-formalists over these potential alternatives. For any formalist agent, define an equivalent anti-formalist agent. This agent shares with the formalist agent the same intersection point (x_c) and *ex-ante* amount of disagreement (A). As shown in the figure, the two agents disagree in the same fraction of cases—indeed the sum of the Areas III and IV is the same. Yet, the equivalent anti-formalist agent always moderates her behavior more (the red triangle is larger than the black triangle) and thus the principal strictly prefers her to the formalist agent.

This section closes with a remark on the value of commitment in claim resolution. Suppose we flipped the order of play and allowed the principal to commit to grant authority in some cases and reverse decisions in other cases. What would she do? Would she be better off with this commitment power. Interestingly, no.

Proposition 5. Suppose the principal could commit to a delegation interval. Facing the anti-formalist agent where $x_c \in [0, 1]$, the principal would deploy the same bounds on discretion as when she could not commit to *ex post* review. Further, she would obtain the same expected payoff.

Proof. For proof, see Appendix. □

Most models, for example Dessein (2002), articulate a substantial difference between delegation and cheap talk games, between commitment and no commitment by the principal. Here, the principal suffers loss from errors alone. He does not obtain a corresponding gain from "correct" decisions. This difference in the utility function from the classic models leads to the result in Proposition 5.

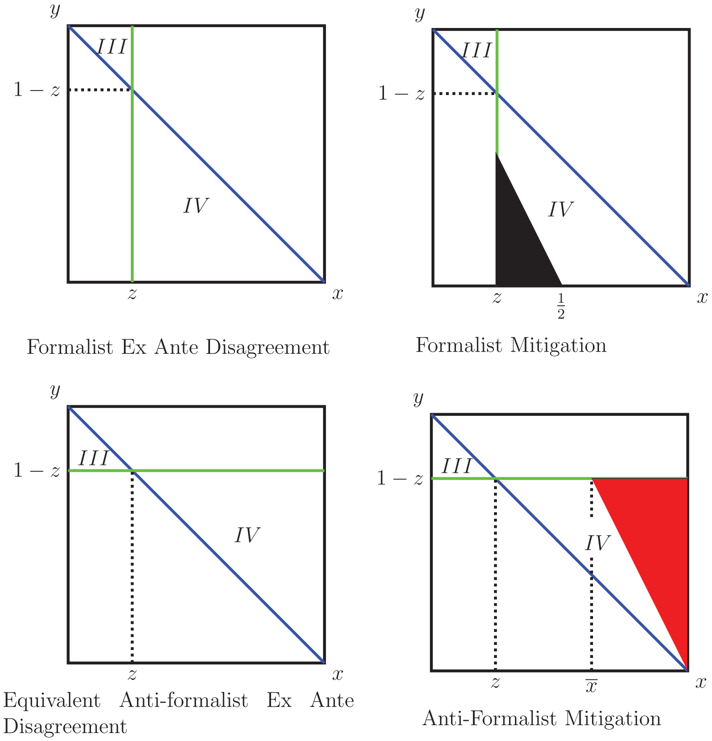
To see why, take a case with a global fact below $\underline{x}$. In the commitment case, the principal suffers a loss from committing to finding this claim invalid. He makes a single type of error: a mistaken finding of invalidity.

In the no-commitment case, the anti-formalist agent partially complies. Upon seeing a valid decision, the principal suffers the same loss from upholding the decision and reversing the decision. In expectation, the principal is equally likely to mistakenly find the claim valid or mistakenly find the claim invalid. And thus while the type of errors shifts in the no-commitment case, the total number of errors remains the same.

Notably, if the principal obtained a gain of, say, 1 from any correct decisions, then she would prefer the setup where the threat of reversal induces some partial compliance by the agent.

5. Discussion

This section identifies key features of our model, discusses the implications they have for our results, and relates them to the most relevant literature.



Formalist Ex Ante Disagreement

Formalist Mitigation

Equivalent Anti-formalist Ex Ante Disagreement

Anti-Formalist Mitigation

Figure 8. Equivalent Formalist and Anti-Formalist Agents.

First, we assume that the principal cannot commit to a review strategy. Typically, the inability to commit transforms delegation games into cheap talk games. In our framework, however, the principal effectively delegates to the agent over an interval in which she affirms the agent's decision with probability one. And that arises even though the principal moves after observing the resolution. The absence of commitment leads to the properties of our equilibrium: outside the interval, the principal reverses unexpected resolutions with some probability less than one.

Second, our model permits two forms of bias: *ex ante* and interim. (In the standard model with constant bias, these two measures collapse into one but in our model they are distinct.) *Ex ante* bias is measured by the percentage A of the claim space on which principal and agent disagree. Interim bias arises after the agent and the principal observe the realization of the global fact and the agent observes the local fact.

The variation in interim bias drives our welfare results. Specifically, whether an agent is interim biased depends on the realization of not just the global fact, but the local fact too. For some realizations, as noted, the agent shares the principal's preferences as to the final resolution.

Third, we characterize formalism or anti-formalism as an inherent trait of the agent. Some judges prefer textual to contextual evidence. Some frontline loan officers prefer to assess an applicant more on how she fares in the interview than the objective markers of financial health. But this is not the only possibility. Judges, for example, might be a formalist/textualist when it comes to contract law and anti-formalist/contextualist in criminal law. In other words, a judge or loan officer could have an inconsistent methodology. Instead of cutlines to partition the claim space, we might have step functions or something else. After all, the weight the agent accords the evidence might itself depend on the global fact. We might reframe our model to be about cases within one field of law (say contract law), but that dodges the central issue: whether a principal would prefer an agent who exhibits a consistent or an inconsistent philosophy as to claim resolution. We leave that question for future work.

Relatedly, what if the only available agents exhibit a specific kind of method disagreement: they are either formalist or anti-formalist. In that case, the welfare claims of the article are less relevant, but the equilibrium predictions of Propositions 1 and 2 remain.

Finally, unlike many auditing models (Andreoni et al. 1998; Cameron et al. 2000), here, the principal cannot pay a cost and observe, at least with some probability, the agent's private information. Suppose we gave our principal that option. For each value of the global fact, she would ask whether the likely error outweighed the cost of investigation. The principal suffers the largest error for cases with global facts located at $\underline{x}$ and $\overline{x}$. Thus, we suspect the principal to be most likely to pay the cost of auditing in these cases. And, as in the conventional inspection game (Fudenberg and Tirole 1991: 17), the principal would randomize between affirming and paying the cost of investigation. The agent would modify her behavior to make this strategy optimal for the principal. Across all global facts, we might observe the principal rely on some combination of investigation and summary reversal of unexpected decisions to induce partial compliance by the agent.

The model most aptly applies where the principal must pay a large cost to observe the local fact, a cost set to infinity in the model. That makes sense, we suspect, for things like the demeanor of the witness or whether a loan applicant appeared to be lying during an interview. The principal cannot "run" the same interview or witness testimony again. The principal might watch a recording, but even that is one step removed and results in a loss of some information. In contrast, the principal can more easily pay an auditing cost to review paper records or documents. Thus, our model more readily applies when the agent observes truly soft information like demeanor, whereas prior work applies to information that is observable, but just at an expense.

6. Conclusion

Debates between formalists and anti-formalists, textualists and contextualists, have been going in the law for a while (Hart et al. 2012; Baude and

Doerfler 2017). What it means to be a formalist or a contextualist has rarely been formalized or subject to a game-theoretic analysis. Our article takes a first step in that direction. In so doing, the model yields two types of results.

First, the two-dimensional structure complicates the ways in which principal and agent may disagree, capturing the core jurisprudential debate. In this framework, the principal generally defers to recommendations of the agent—but does not need to commit to do so. The nature of the delegation depends on the nature of the disagreement between principal and agent. Anti-formalist agents whose cutlines intersect the principal's cutline in the open interval (0, 1) have two bounds on discretion. Their decisions at extreme values of the global fact x are subject to review; consequently, the principal might overrule holdings of both validity and invalidity. For formalist agents, overruling can only occur for intermediate values of x.

Second, we show that the principal has preferences over the bias that infects her agent. Conditional on a fixed level of *ex ante* disagreement, the principal prefers an anti-formalist agent to any other similar-situated agent.

While formalism only brings costs, anti-formalism brings costs and benefits to the principal. Anti-formalism implies that any decision will reflect a hefty dose of private information; and private information is valuable when the global fact is uninformative. But anti-formalism also implies that the agent has a tendency to disregard, or at least underweight, the information also available to the principal. This disregard imposes costs on the principal in those cases in which she finds the global fact highly informative. Yet, the principal can partially temper the agent's anti-formalism in these contexts by credibly threatening, conditionally on the realized global fact, to reverse some unexpected decisions. And this credible threat, in turn, induces some—albeit imperfect—compliance by the anti-formalist agent. This formalist agent, in contrast, is much harder to motivate.

Our model has several applications. Delegation is widespread in both private and public bureaucracies. Our first result, however, has special leverage on delegation in the public sphere. Generally, in the public sector, the principal cannot commit to her delegation through an enforceable contract. In public bureaucracies, this inability to commit derives from the limitations on the employment contract. In the federal judiciary, though a hierarchy of courts exists, there are no mechanisms of control other than affirmance and reversal of the decisions of the lower court. An appellate court, thus, cannot commit to defer to the decisions of a lower court or an administrative agency. Our analysis shows that, nonetheless, when the agent has private information that is valuable to her, the principal will rationally delegate many classes of decision problems to the agent.

We conclude with two suggestions for future work. First, the model assumes that the two facts are independently distributed. The principal does not learn anything about the location of the local fact from the realization of the global fact. If the two facts were perfectly correlated, of course, the principal would not need the agent at all. Partial correlation might allow the principal to do better by improving the amount of partial compliance by the agent. But, we suspect, the driving force behind the desirability of anti-formalist agents would remain.

Unlike the formalist, the anti-formalist agent disagrees as to what claims should be found valid in cases where the global fact points toward invalidity, making threats of reversal credible. The independence of the draws does not dictate the economic insight. Instead, the driving force is that anti-formalist agents have cutlines with a flatter slope than the principal while formalists have cutlines with a steeper slope.

Second, the model ignores the decision of whether to file a claim in the first place. In the context of litigation, the plaintiff might not file if she knows, says, that a finding of liability is unlikely to arise. This changes the kinds of cases the principal and agent potentially consider, which, in turn, would surely change the equilibrium strategies.

Appendix

A. Uninformative Equilibria

There are two ways in which the agent's decision might fail to convey information about the local fact. First, irrespective of type, the agent might, for example, send the message "valid" and "invalid" with equal probability. Alternatively, the agent might pool on the expected message, the message which accords with the global fact. We consider each type of uninformative equilibria in turn.

First, suppose $d = 1/2$ for all cases. In that setting, the principal's beliefs about y remain uniform with support $[0, 1]$ for all cases and all agent decisions. And so, the principal's best response is

$$\gamma^\star = g^\star(x, d) = \begin{cases} 1 & \text{if } x < \dfrac{1}{2} \text{ and } d = 1 \\ 1 & \text{if } x > \dfrac{1}{2} \text{ and } d = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Take the agent type where $y = 1$ who observes a global fact above 1/2. If she follows this babbling strategy, her expected payoff is -1/2. In contrast, her payoff to sending the valid message is 0. This agent type thus has a profitable deviation, meaning this equilibrium cannot exist.

Second, let us examine a pooling equilibrium. In this equilibrium, the agent decides on the basis of the global fact alone. If the global fact exceeds 1/2, she finds valid; if it is less than 1/2, the agent decides invalid.

Take a case where $x > 1/2$. Suppose all agent types decide the claim as valid. Off the equilibrium path, the principal believes that the message "invalid" is equally likely to come from any agent type. Given the agent's behavior and the principal's beliefs, the principal affirms the valid decision and reverses the invalid decision. No agent type wants to deviate and incur the cost of reversal. The pooling equilibrium therefore exists. The issue is whether the principal's beliefs supporting this equilibrium are plausible. We next demonstrate that the beliefs fail to be "universally divine" as defined by Banks and Sobel (1987).

First, for agents where $y > f(x)$, the decision "valid" provides a higher payoff than the decision "invalid" irrespective of whether the principal affirms or reverses the invalid decision. Her equilibrium payoff is 0. If this agent type reports "invalid" either (a) the principal reverses and the agent suffers a reversal cost or (b) the principal affirms the invalid decision and this agent suffers from a mistaken final resolution. Thus, any equilibrium cannot have the principal believe the invalid message came from a type in the interval $y \in [f(x), 1]$.

For the remaining types, define $\gamma(y)$ as the probability of reversal of an invalid decision such that the agent who draws y is indifferent between sending the message valid and invalid; that is, $\gamma(y)(1 + ky) = 1$ or $\gamma(y) = \frac{1}{1+ky}$.

Following the proof in Reinganum and Wilde (1986) observe that $\gamma(y)$ is maximized at $y = 0$. Thus, the agent who draws $y = 0$ is most likely to deviate from sending the message valid (that is, she deviates for the most values of $\gamma(y)$). Universal divinity, therefore, demands the principal believe that the invalid message came from this agent, the one who drew $y = 0$. Given those beliefs, the principal wants to affirm the invalid decision. Anticipating this affirmance, any agent who drew a $y < f(x)$ would prefer to deviate and send the invalid message rather than pool on the valid message. And so, the pooling equilibrium fails to be universally divine.

B. Alternative Payoff Function

This subsection shows that the equilibrium described in Proposition 1 is robust to a specification of preferences where the reversal cost does not vary with the local fact. Define the utility of the principal as

$$U_p(x, y, r) = -r(1 - x - y)I_{y<1-x} - (1 - r)(1 - I_{y<1-x})(y - (1 - x)).$$

Define the utility of the agent as

$$U_a(x, y, r, d, \gamma) = -rI_{y<f(x)}(f(x) - y)) - (1 - r)(1 - I_{y<f(x)})(y - f(x)) - \gamma k.$$

where k is a constant. With these utility functions, the players suffer greater disutility when the mistake in the resolution is big rather than small. Further, the agent suffers a fixed cost of reversal, which is independent of her type.

Assume that $w < \frac{1}{2}$. With these utility functions, Equations (1) and (2) become

$$-\frac{\int_{\max\{0,f(x)\}}^{1-x}(1-x-y)\mathrm{d}y}{\mathrm{pr(valid)}} = -\min\left\{\frac{(1-x)^2}{2\,\mathrm{pr(valid)}}, \frac{(1-x-f(x))^2}{2\mathrm{pr(valid)}}\right\}.$$

And

$$-\frac{\int_{1-x}^{1}(y-(1-x))\mathrm{d}y}{\mathrm{pr(valid)}} = -\frac{x^2}{2\,\mathrm{pr(valid)}}.$$

The principal will affirm any valid decision if $x > \underline{x}$ where $\underline{x}$ solves

$$-\min\left\{\frac{(1-x)^2}{2\,\mathrm{pr(valid)}}, \frac{(1-x-f(x))^2}{2\,\mathrm{pr(valid)}}\right\} + \frac{x^2}{\mathrm{pr(valid)}} = 0.$$

One solution is $\underline{x} = \min\left\{\frac{1}{2}, \frac{1-w-z}{2-3w}\right\}$, which is the same as Equation (4). For claims less than $\underline{x}$, the equilibrium is defined as the joint solution to an equation making the principal indifferent between reversing and affirming a valid decision and an equation ensuring that the reversal probability induces that behavior by the agent. That is,

$$-\frac{\int_{y^\star}^{1-x}(1-x-y)\mathrm{d}y}{\mathrm{pr(valid)}} + \frac{\int_{1-x}^{1}(y-(1-x)\mathrm{d}y}{\mathrm{pr(valid)}} = 0.$$

And

$$-(y-f(x)) + \gamma \times (y - f(x) + k) = 0$$

The principal's mixing condition reduces to

$$-\frac{(1-x-y^\star)^2}{2\,\mathrm{pr(valid)}} + \frac{x^2}{2\,\mathrm{pr(valid)}} = 0.$$

Considering only values of $y^\star < 1$, the equation admits the positive solution:

$$y^\star = 1 - 2x.$$

Using this, the reversal probability is

$$\gamma^\star = \frac{1-2x-f(x)}{1-2x-f(x)+k}.$$

Turn next to invalid decisions. The principal's payoff from affirming is

$$-\frac{\int_{1-x}^{\min\{1,f(x)\}}(y-(1-x))\mathrm{d}y}{\mathrm{pr(invalid)}}, \tag{A1}$$

while the payoff to reversing is

$$-\frac{\int_{0}^{1-x}(1-x-y)\mathrm{d}y}{\mathrm{pr(invalid)}}. \tag{A2}$$

The global fact, where Equation (A1) equals Equation (A2), defines $\overline{x}$. Doing the integration and solving for one (of two) solutions, we get

$$\overline{x}=\max\left\{\frac{1}{2},\frac{2-z-2w}{2-3w}\right\}.$$

For cases where $x>\overline{x}$, the principal must be willing to mix between affirming and reversing an invalid decision given the agent plays $y^{\star}$. Thus,

$$\begin{aligned}0&=-\frac{\int_{1-x}^{y\star}(y-(1-x))\mathrm{d}y}{\mathrm{pr(invalid)}}+\frac{\int_{0}^{1-x}(1-x-y)\mathrm{d}y}{\mathrm{pr(invalid)}}\\&=-\frac{(y^{\star}-(1-x))^2}{2\ \mathrm{pr(invalid)}}+\frac{(1-x)^2}{2\,\mathrm{pr(invalid)}}\end{aligned}$$

from which we derive the positive solution $y^{\star}=2(1-x)$. Meanwhile, for the agent to prefer the cutoff point $y^{\star}$ demands that reversal probability make her indifferent at that value. That is,

$$(f(x)-y^{\star})-\gamma\times(f(x)-y^{\star}+k)=0$$

or,

$$\gamma^{\star}=\frac{f(x)-2(1-x)}{f(x)-2(1-x)+k}.$$

And thus, we see that the results from Proposition 1 are robust to alternative specifications of the utility functions.

C. Proof of Proposition 3

In this proof, we construct a three-partition equilibrium for global facts between $[0,x_c]$. The proof for global facts between $[x_c,1]$ is similar.

Consider first a global fact in the interval $[\underline{x},x_c]$. Partition the space of local facts into three intervals, defined by $[0,y_1],[y_1,y_2]$ and $[y_2,1]$. Suppose the agent sends message invalid$_1$ for local facts in the first interval; invalid$_2$ for local facts in the second interval, and valid for local facts in the third interval. Set $y_2=f(x)$. These messages induce uniformly distributed beliefs by the principal with the support defined by the length of each interval. And so, the principal will find the claim invalid when she sees invalid$_1$ or invalid$_2$. She will also affirm and find the claim valid when she sees the valid message since $x \in [\underline{x}, x_c]$. In equilibrium, the principal suffers the same expected loss as with just two messages; that is, $1 - x - f(x)$.

Consider next a global fact where $x < \underline{x}$. Again, partition the space of local facts into three intervals. $[0, y_1], [y_1, y_2]$ and $[y_2, 1]$. Let $y_2 = 1 - 2x$. Again, take three messages: (a) invalid$_1$; (b) invalid$_2$; and (c) valid. Given the y_2, the principal is willing to mix between reversing and not following a valid message (i.e., issuing a final ruling of valid or invalid). Suppose she mixes with $\gamma^\star = \frac{1}{1+2kx}$. Given her beliefs, the principal finds the claim invalid if she observes invalid$_1$ or invalid$_2$.

No agent who draws a local fact in the first interval has an incentive to deviate. If she reports invalid$_2$, the principal finds the claim invalid, which does not improve her payoff. If the agent deviates and reports valid, the principal reverses with probability $\gamma^\star$, leading to a payoff of

$$-\gamma^\star(1 + k(1 - y)).$$

which is less than –1 for all $y < 1 - 2x$. The same analysis applies to agents who draw a local fact in the second interval. Finally, no agent who draws a local fact in the interval between $[1 - 2x, 1]$ has an incentive to deviate. If they do, the principal will resolve the claim as invalid, leading to a lower payoff.

This three-partition equilibrium provides the exact same expected payoff for the principal as the two-step partition derived in Proposition 1. The principal suffers a loss, in expectation, of x, for global facts in the range $[0, \underline{x})$.

D. Solution to Constrained Optimization Problem

Recall that the principal maximizes

$$W = -\frac{(1 - w - z)^2 + (z - w)^2}{2(2 - 3w)(1 - 2w)}. \tag{A3}$$

subject to

$$\frac{(1 - w - z)^2 + (z - w)^2}{2(1 - w)(1 - 2w)} - A = 0.$$

The *ex ante* disagreement constraint can be expressed as

$$(1 - w - z)^2 + (z - w)^2 = 2A(1 - w)(1 - 2w). \tag{A4}$$

Substitute the RHS of Equation (A4) into the numerator of Equation (A3). Doing so eliminates z and transforms the constrained problem into an unconstrained one. Some cancellations reveal the principal's program as

$$\max_{w} -\frac{A(1-w)}{2-3w}. \tag{A5}$$

Figure 9 is the graph of $W(w)$ where $A = 1/4$. It shows that W decreases with w. All else equal, the principal prefers to set $w = 0$. Yet any solution must ultimately involve a real number for z. To account for this fact, solve the constraint for z.

$$z = \frac{1 \pm \sqrt{8Aw^2 - 4w^2 + 4w - 12Aw + 4A - 1}}{2} \tag{A6}$$

To ensure z is real, the expression under the square root must be positive, or

$$8Aw^2 - 4w^2 + 4w - 12Aw + 4A - 1 \geq 0 \tag{A7}$$

The LHS of Equation (A7) has two roots: $\underline{w} = \frac{1-4A}{2-4A}$ and $1/2$. Further, the expression is only positive for values of w in the interval $[\underline{w}, \frac{1}{2}]$. Finally, notice that $\underline{w}$ is positive if $A < 1/4$ and negative if $A > 1/4$. In short, the values of w where the expression under the square root is (strictly) positive, includes 0 when A exceeds $1/4$. Otherwise it does not.

Because the principal's welfare decreases in w, she wants w to be as small as possible. The assumption of limited disagreement means that $A < 1/4$. Thus, the smallest available selection for w is $\underline{w} > 0$. At this value, we also have that $x_c = \frac{1-w-x}{1-2w} \in (0,1)$ as required by Assumption 1.

On the other hand, if $A \geq 1/4$, the principal is free to set $w = 0$.

Finally, if the solution to the problem involves setting $w = \underline{w}$ then by Equation (A6), we get $z = 1/2$. If the solution is $w = 0$ then by Equation (A6) we have $z = \frac{1 \pm \sqrt{4A-1}}{2}$.

E. Proof of Proposition 4

i. Part 1. Using the optimal anti-formalist agent, the principal's welfare is

$$W^{\star}(A) = -\frac{A}{1+4A}.$$

Consider an agent who is lax, but agrees about the method $(w = \frac{1}{2}, z < \frac{1}{2})$ (The proof for the strict agent is similar). The area of disagreement between the principal and this agent is

$$\frac{1}{2} - \frac{(2z)^2}{2}$$

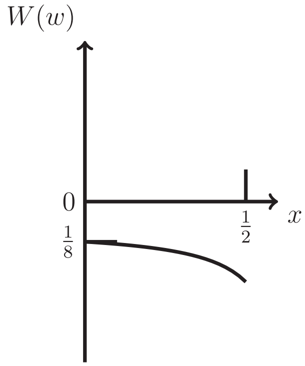


Figure 9. Graph of $W(w)$.

Set this area equal to A and solve the constraint for positive value z.

$$z(A) = \frac{\sqrt{1-2A}}{2}. \tag{A8}$$

Observe that $z(A)$ decreases in A. Further, at $A = 1/4$, Equation (A8) provides $z = \sqrt{2}/4$. We next use the assumption about the extent of disagreement to restrict the parameters (w, z) under consideration,

Assumption 1 restricts attention to $A \leq 1/4$. And thus, we only examine values of z in the interval $\left[\frac{\sqrt{2}}{4}, \frac{1}{2}\right]$.

Recall that $\underline{x}$ is

$$\underline{x} = \min\left\{\frac{1}{2}, \frac{1-w-z}{2-3w}\right\}.$$

We assume that this agent agrees with the principal as to the threshold; that is, $w = 1/2$. Plugging in $w = 1/2$ into $\underline{x}$ yields a lower bound of $1 - 2z$.[9] Facing this agent, the principal's welfare is the *ex ante* area of disagreement less the area of mitigation, or

$$\begin{aligned} W_t(A) &= -\left(A - \frac{\underline{x}(1-2z)}{2}\right) \\ &= -\left(A - \frac{(1-2z)^2}{2}\right). \end{aligned} \tag{A9}$$

Plug the value of z from Equation (A8) into Equation (A9). Doing so, we get

$$W_t(A) = (1-2A) - \sqrt{1-2A}$$

which is less than $W^\star(A) = -\frac{A}{1+4A}$ for all $A \in (0, \frac{1}{4}]$.

[9] We know that $1 - 2z < \frac{1}{2}$ since $z > \frac{\sqrt{2}}{4} > \frac{1}{4}$.

ii. Part 2. Start with a formalist agent who disagrees in A cases and does not face a reversal threat. The principal's payoff from employing this agent is—A, which is less than $W^{\star}(A)$.

Next recall that the formalist agent can never face a reversal threat for both valid and invalid decisions. Picking one, we focus on a formalist agent who faces a reversal threat for some valid decisions, and therefore partially complies where she prefers valid and the principal prefers invalid. This means we consider the case where $z \in \left[0, \frac{1}{2}\right]$

Our first step is to show that we only need to consider formalist agents where $w \in (\frac{2}{3}, 1]$ given our assumptions.

First, Assumption 1 restricts attention to $x_c = \frac{w+z-1}{2w-1} \in (0, 1)$. This implies

$$z > 1 - w \tag{A10}$$

For the formalist agent to moderate her behavior with respect to validity requires

$$f(x) < 1 - 2x \tag{A11}$$

for a value of $x \in [0, \frac{1}{2}]$. Solving Equation (A11) as an equality gives

$$\underline{x} = \frac{z + w - 1}{3w - 2} \tag{A12}$$

Note that Equation (A12) is less than 1/2 when

$$z < \frac{w}{2}. \tag{A13}$$

Combining the inequalities in Equations (A10) and (A13) yields

$$1 - w < z < \frac{w}{2},$$

Ensuring that a z exists in this range restricts the formalist agents under consideration: it must be that $w > 2/3$.

Given the analysis thus far and the attention on formalist agents who face reversal threats as to valid decisions, the remainder of the proof only considers agents where $w > 2/3$ and $z < 1/2$. Such an agent partially complies for cases with global facts in the interval $[\underline{x}, \frac{1}{2}]$.

This formalist agent *ex ante* disagrees in the following percentage of cases.

$$\frac{(z + w - 1)^2 + (w - z)^2}{2(2w - 1)w}.$$

The principal's welfare from hiring this agent is the *ex ante* disagreement less than benefits of mitigation.

$$W_f(w,z) = -A + \frac{(w-2z)^2}{4w(3w-2)}.$$

Define an "equivalent" anti-formalist agent by the pair $(\tilde{w}, \tilde{z})$, where

$$\tilde{w} = 1 - w$$
$$\tilde{z} = 1 - z.$$

The disagreement area associated with the equivalent anti-formalist agent is

$$\begin{aligned} A &= \frac{(1-\tilde{w}-\tilde{z})^2 + (\tilde{z}-\tilde{w})^2}{2(1-\tilde{w})(1-2\tilde{w})} \\ &= \frac{(z+w-1)^2 + (w-z)^2}{2w(2w-1)}, \end{aligned}$$

which is the same as the formalist agent. Likewise, the point at which the cutlines cross is

$$\begin{aligned} x_c &= \frac{1-\tilde{w}-\tilde{z}}{1-2\tilde{w}} \\ &= \frac{w+z-1}{2w-1}, \end{aligned}$$

which is the same as the formalist agent.

The next step is to show that the principal achieves a higher welfare from employing the "equivalent" anti-formalist agent than the formalist counterpart. As a result, among similarly situated agents—those that disagree in A cases—the principal can always do better by hiring the equivalent anti-formalist.

The welfare associated with hiring the equivalent anti-formalist agent is

$$\begin{aligned} W_e(\tilde{w},\tilde{z}) &= -A + \frac{(1-\tilde{w}-\tilde{z})^2}{2(1-\tilde{w})(2-3\tilde{w})} + \frac{(\tilde{z}-\tilde{w})^2}{2(2-3\tilde{w})(1-\tilde{w})} \\ &= -A + \frac{(w+z-1)^2}{2w(3w-1)} + \frac{(w-z)^2}{2w(3w-1)}. \end{aligned}$$

Observe that $W_e(\tilde{w},\tilde{z}) > W_f(w,z)$ if

$$\frac{(w-z)^2}{2w(3w-1)} > \frac{(w-2z)^2}{4w(3w-2)}.$$

Let the difference between these expressions be

$$D(w,z) = \frac{(w-z)^2}{2w(3w-1)} - \frac{(w-2z)^2}{4w(3w-2)}$$
$$= \frac{3w^2 - 3w - 6z^2 + 4z}{4(3w-1)(3w-2)}$$

Observe in Figure 10 that over the relevant range of w, $D(w, z)$ increases with w. Moreover, the function equals zero at

$$\hat{w}(z) = \frac{3 \pm \sqrt{9 - 48z + 72z^2}}{6}$$

Focusing on the larger root, we next show that $D(w, z)$ must always be positive. That amounts to showing that $w > \hat{w}$ when $w \in \left[\frac{2}{3}, 1\right]$ and $z \in \left[0, \frac{1}{2}\right]$

Define

$$G(z) := 1 - z - \hat{w}(z)$$
$$H(z) := 2z - \hat{w}(z)$$

Recognizing a few facts about $G(z)$ and $H(z)$ finishes the proof.

(1) $G(0) = 0$, $G\left(\frac{1}{3}\right) = 0$.

(2) We have

$$G'(z) = \frac{4 - 12z}{\sqrt{9 - 48z + 72z^2}} - 1$$
$$G''(z) = -\frac{4}{(24z^2 - 16z + 3)\sqrt{9 - 48z + 72z^2}} < 0$$

Using the above expressions, notice that $G(z)$ has a single critical point at $z = 0.21$, which is the maximum. Combined with the facts that $G(0) = 0$ and $G\left(\frac{1}{3}\right) = 0$ observe that $G(z) > 0$ when $z \in (0, \frac{1}{3})$.

(3) $H\left(\frac{1}{3}\right) = 0$, $H\left(\frac{1}{2}\right) = \frac{1}{2} - \frac{\sqrt{3}}{3} > 0$

(4) We also have that $H'(z) = 2 + \frac{4-12z}{\sqrt{9-48z+72z^2}} > 0$ over the relevant range of z. As a result $H(z) \geq 0$ if $z \in \left(\frac{1}{3}, \frac{1}{2}\right]$.

To satisfy the restriction on the parameters, Equation (A10) demands that $w > 1 - z$. It follows that

$$w - \hat{w} > 1 - z - \hat{w} = G(z) \geq 0.$$

Since (a) w is greater $\hat{w}$ and (b) $D(w, z)$ is positive when $w > \hat{w}$, we get $D(w,z) > 0$ when the agent's cutline is defined by $w \in \left(\frac{2}{3}, 1\right]$ and $z \in [0, \frac{1}{3}]$.

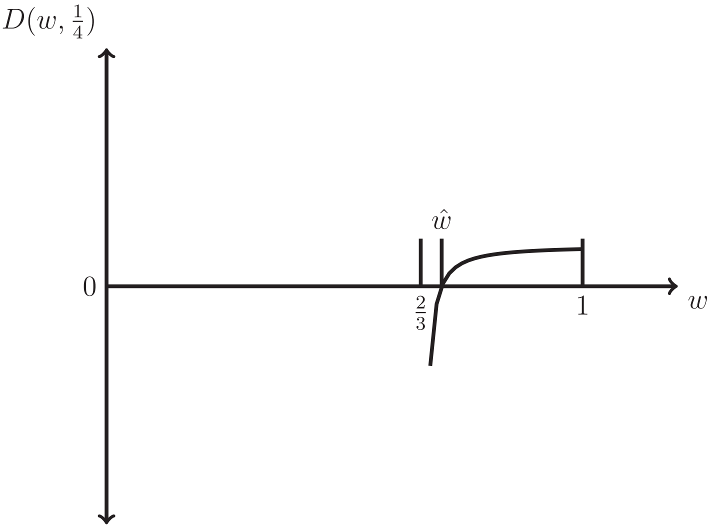


Figure 10. Graph of $D(w, z)$ where $z = \frac{1}{4}$.

We next have one last formalist agent to consider. This agent's cutline is defined by $w \in \left(\frac{2}{3}, 1\right]$ and $z \in \left(\frac{1}{3}, \frac{1}{2}\right]$.

Restriction Equation (A13) requires that $w > 2z$. It follows that

$$w - \hat{w} > 2z - \hat{w} = H(z) \geq 0.$$

Therefore $w > \hat{w}$ and $D(w, z) > 0$ in this case.

To close, because $D(w, z) > 0$ for all formalist agents whose preferences are consistent with Assumption 1, it follows that $W_e(\tilde{w}, \tilde{z}) > W_f(w, z)$. And, of course, $W^\star(A) > W_e(\tilde{w}, \tilde{z})$, completing the proof.

F. Proof of Proposition 5

Facing an anti-formalist agent, the principal who can commit selects a lower bound, $\underline{x}$ and an upper bound $\overline{x}$ to maximize

$$W = -\Bigg\{ \int_0^{\underline{x}} x dx + \int_{\underline{x}}^{x_c} (1 - x - \left(\frac{z}{1-w} - \frac{wx}{1-w}\right) dx + \int_{x_c}^{\overline{x}} \left(\frac{z}{1-w} - \frac{wx}{1-w} - (1-x)\right) dx + \int_{\overline{x}}^{1} (1-x) dx \Bigg\}.$$

The solution is familiar. It is $\underline{x} = \frac{1-w-z}{2-3w}$ and $\overline{x} = \frac{2-2w-z}{2-3w}$. Moreover, the value of the objective function is the same as without commitment.

References

Aghion, P., and J. Tirole. 1997. "Formal and Real Authority in Organizations," 105 *Journal of Political Economy* 1–29.

Alonso, R., and N. Matouschek. 2008. "Optimal Delegation," 75 *The Review of Economic Studies*, 259–93.

Andreoni, J., B. Erard, and J. Feinstein. 1998. "Tax Compliance," 36 *Journal of Economic Literature* 818–60.

Banks, J. S., and J. Sobel. 1987. "Equilibrium Selection in Signaling Games," 55 *Econometrica* 647–61.

Baude, W., and R. D. Doerfler. 2017 "The (Not so) Plain Meaning Rule," 84 *The University of Chicago Law Review* 539–566.

Bendor, J., A. Glazer, and T. Hammond. 2001. "Theories of Delegation," 4 *Annual Review of Political Science* 235–69.

Bernstein, L. 2015. "Custom in the Courts," 110 *Northwestern University Law Review 63–113*.

Bueno de Mesquita, E. and M. Stephenson. 2002. Informative Precedent and Intrajudicial Communication," 96 *American Political Science Review* 755–66.

Cameron, C. M., J. A. Segal, and D. Songer. 2000. "Strategic Auditing in a Political Hierarchy: An Informational Model of the Supreme Court's Certiorari Decisions," 94 *American Political Science Review* 101–16.

Charny, D. 1999. "The New Formalism in Contract," 66 *The University of Chicago Law Review* 842–57.

Che, Y., and N. Kartik. 2009. "Opinions as Incentives," 117 *Journal of Political Economy* 815–60.

Corbin, A. L. 1964–1965. "Interpretation of Words and the Parol Evidence Rule," 50 *Cornell Law Quarterly* 161–190.

Crawford, V. P., and J. Sobel. 1982. "Strategic Information Transmission," 50 *Econometrica* 1431–51.

Dessein, W. 2002. "Authority and Communication in Organizations," 69 *Review of Economic Studies* 811–38.

Dewatripont, M., and J. Tirole. 1999. "Advocates," 107 *Journal of Political Economy* 1–39.

Epstein, D., and S. O'Halloran. 1999. *Delegating Powers: A Transaction Cost Politics Approach to Policy Making under Separate Powers*. 1st ed. Cambridge: Cambridge University Press.

Fudenberg, D. and J. Tirole. 1991. *Game Theory*. 1st ed. Cambridge, MA: The MIT Press.

Gennaioli, N. and A. Shleifer. 2007. "The Evolution of Common Law," 115 *The Journal of Political Economy* 43–68.

Godbillon-Camus, B. and C. J. Godlewski. 2013. "Risk Management, Soft Information and Bankers' Incentives," 123 *Revue D'conomie Politique* 763–91.

Goetz, C. J. and R. E. Scott. 1985. "The Limits of Expanded Choice: An Analysis of the Interactions between Express and Implied Contract Terms," 73 *California Law Review* 261–322.

Hart, H. L. A., L. Green, J. Raz, and P. A. Bulloch. 2012. *The Concept of Law*. 3rd ed. Oxford: Oxford University Press.

Holmstrom, B. 1984. "On the Theory of Delegation," in M. Boyer and R. E. Kihlstrom, eds., *Bayesian Models in Economic Theory*, 115–41. Amsterdam, the Netherlands: North-Holland.

Lax, J. R. 2012. "Political Constraints on Legal Doctrine: How Hierarchy Shapes the Law," 74 *The Journal of Politics* 765–81.

Liberti, J. M., and A. R. Mian. 2009. "Estimating the Effect of Hierarchies on Information Use," 22 *The Review of Financial Studies* 4057–90.

Liberti, J. M., and M. A. Petersen. 2019. "Information: Hard and Soft," 8 *The Review of Corporate Finance Studies* 1–41.

Manuel, A., and K. Bagwell. 2013. "The Theory of Optimal Delegation with an Application to Tariff Caps," 81 *Econometrica* 1541–99.

Reinganum, J. F., and L. L. Wilde. 1986. "Settlement, Litigation, and the Allocation of Litigation Costs," 17 *The RAND Journal of Economics* 557–66.

Scott, R. E. 1999. "The Case for Formalism in Relational Contract Symposium in Honor of Ian R. MacNeil: Relational Contract Theory: Unanswered Questions," 94 *Northwestern University Law Review* 847–76.

Spence, M. 1973. "Job Market Signaling," 87 *Quarterly Journal of Economics* 355–374.