I Don't Know*

Matthew Backus† Andrew T. Little‡

May 14, 2018

Abstract

Experts with reputational concerns, even good ones, are averse to admitting what they don't know. This diminishes our trust in experts and, in turn, the role of science in society. We model the strategic communication of uncertainty, allowing that some questions are unanswerable. Combined with a new use of Markov sequential equilibrium, we recast prior negative results in a simple and stark light, and thereby obtain elusive positive results. When problems are potentially unanswerable, checking features of the problem itself that good experts will infer is more useful for inducing honest reporting than checking predictive success.

*Thanks to Charles Angelucci, Jonathan Bendor, Wouter Dessein, James Hollyer, Navin Kartik, Greg Martin, Andrea Prat, Daniel Rappaport, Jim Snyder, Joel Sobel, several anonymous referees and audiences at MPSA 2015, EARIE 2017, The 28th International Game Theory Conference, QPEC 2017, Petralia Workshop 2017, SAET 2017, Columbia, Harvard, the Higher School of Economics, Peking University, and Stanford for thoughtful comments and suggestions. We thank Alphonse Simon and Brenden Eum for excellent research assistance. All remaining errors are our own.

†Columbia and NBER, matthew.backus@columbia.edu

‡UC Berkeley, andrew.little@berkeley.edu

> [...] it is in the admission of ignorance and the admission of uncertainty that there is a hope for the continuous motion of human beings in some direction that doesn't get confined, permanently blocked, as it has so many times before in various periods in the history of man.
>
> — Richard Feynman, John Danz Lecture, 1963

> It seemed to him that a big part of a consultant's job was to feign total certainty about uncertain things. In a job interview with McKinsey, they told him that he was not certain enough in his opinions. "And I said it was because I wasn't certain. And they said, 'We're billing clients five hundred grand a year, so you have to be sure of what you are saying.'"
>
> — Michael Lewis quoting Daryl Morey in *The Undoing Project*, 2016

1 Introduction

Executives are experts in the domain of decision-making, but not experts in the domains in which they make decisions. As a result, organizations are built around them to aggregate expertise – they hire consultants, employ specialists, and take the testimony of experts in order to make better choices.[1] However, a large literature following in the wake of Crawford and Sobel (1982) has made clear that experts may have an incentive to distort advice, pander to the decision-makers beliefs, and overstate the precision of their information. This limits the information transmitted in equilibrium and leads to worse choices, as decision-makers are rationally unmoved by strategic advice.

Here we focus on particularly prickly part of this problem. Experts frequently don't know the answer to questions they are asked. Their ignorance can be driven by a lack of knowledge, or the fact that some questions truly are unanswerable, e.g. predicting close elections or making price recommendations when demand is unidentified in the available data. Since incompetent experts are more likely to not have the answers, admitting uncertainty comes with a reputational cost. In order to prevent uninformed experts from feigning knowledge and pushing executives to risky decisions, they must be willing to say "I don't know".

[1] See Lazear (2005) for the original formulation of this argument with respect to CEOs.

Can experts who care about perceptions of their competence be induced to admit uncertainty? On this the prior literature is particularly bleak. We contribute by introducing a cheap talk model with an explicit focus on problem difficulty, and showing that it creates room for positive results. Consistent with prior work, fact-checking experts – with the attendant threat of reputational consequences when wrong – is never enough to induce honesty. In the language of our model, they never say "I don't know." However, new to our setting, we show that if the decision-maker can learn, *ex-post*, whether the question was well-formulated, then the threat of catching the expert answering an unanswerable question can make experts willing to admit uncertainty.

A direct and timely implication of our finding concerns the management of experts in organizations. Should they be assigned to project teams, evaluated by the objective outcomes of A/B tests, or should they be managed by other experts in semi-independent "research labs"? Our results suggests that, particularly in the domain of difficult questions, the most effective way to discipline reputational concerns among experts may be to have them managed by other experts. This is consistent with the recent trend of hiring academic economists to lead research labs in the technology sector, an environment with an abundance of new and challenging questions; some well-posed, and some less so.

The main innovation of the model is an explicit focus on heterogeneity in the *difficulty* of problems. We view this as a salient and neglected source of uncertainty for real decision-makers: often, because of the same inexpertise that requires them to hire experts whose quality they cannot evaluate, they ask questions that cannot be answered.[2] Where the canonical principal-expert framework focuses on epistemic uncertainty ("what is the state of the world?"), the notion of problem difficulty introduces aleatory uncertainty ("is the state of the world knowable?"). We believe that economists who have worked with policy-makers or firms will find this notion familiar.[3]

To make our contribution clear, we introduce the concept of problem difficulty into a principal-expert model in the simplest possible fashion. The state of the world, expert quality, and problem difficulty are all binary. Bad experts learn nothing about the state of the world. Good experts learn the state if and only if the problem is solvable. All ex-

[2]That a question cannot be answered does not imply that the answer does not exist. An economic consultant may be asked to estimate demand using data in which it is not econometrically identified, or a political pundit might be asked to predict the outcome of an election. In both cases, an answer exists and may ultimately be revealed, whether by price experiments or simply waiting.

[3]Although prior work has allowed for varying precision of expert information, the key difference here is that difficulty is a property of the problem itself.

perts then communicate a message to the decision-maker, who takes an action. Finally, the decision-maker – potentially endowed, *ex post*, with information about the state or problem difficulty – forms posterior beliefs about the expert quality. In conjectured honest equilibria, in which experts reveal their decision-relevant information, off-path beliefs embed threats: if I catch you red-handed in a lie, I will believe that you are a bad expert. But are such beliefs credible? They are off-path, and so we need more discipline than perfect Bayesian equilibrium offers – in particular, to account for the simplicity of our binary setup, we want an equilibrium notion that imposes a notion of robustness on off-path beliefs.

Therefore we study Markov sequential equilibria, a solution concept that combines the beliefs of sequential equilibrium (Kreps and Wilson, 1982) with the restriction to Markov strategies.[4] Reputational games are a particularly natural setting in which to study Markov sequential equilibrium because in such games, good behavior is induced by the threat of bad beliefs. But threatened beliefs, just like ordinary threats (Selten, 1978), may not be credible. Markov sequential equilibrium requires that the threatened beliefs be structurally sound, in the following sense: first, that they be derived from a feasible strategy given the information structure of the game and second, that they do not update on payoff-irrelevant information. This restriction sheds light on the simple logic of the negative results concerning honesty that precede us in the literature: facing an unanswerable question, good experts and bad experts are strategically equivalent, and so the severity of our beliefs is bounded by the conflation. Therefore, uninformed experts can always improve their lot by guessing, and so honesty is never an equilibrium strategy.

Negative results concerning honesty in reputational games are neither new nor surprising. Rather, our main contributions are twofold. First, to our knowledge ours is the first model to formulate and exploit the restriction imposed by Markov sequential equilibrium on off-path beliefs. In section 5 we show how this casts new light on the structure of the reputational problem. Second, guided by that insight, we offer new positive results concerning the feasibility of admitting uncertainty. The restriction imposed my MSE makes clear the symmetry that we need to break in order to induce honesty: the principal needs to learn something that allows them to differentially affect the incentives of good experts facing an impossible question and bad experts. Validating problem difficulty does precisely this –

[4]We will be precise in Section 4, but the novel restriction of Markov sequential equilibrium is that off-path beliefs must be rationalizable as the limit of a sequence of Markov strategies. Looking for perfect Bayesian equilibrium or sequential equilibrium in this setting yields what we consider to be unreasonable equilibria that violate this restriction, although many of the qualitative features of the results still hold. The interested reader can anticipate a discussion of this at the end of Section 6.

because it is a property of the question itself, and because it is observable to good experts. We believe this mechanism may be more general, suggesting an avenue for future work. Though we have focused on problem difficulty here, in general, any feature of the problem that the good experts are more likely to observe creates an informational asymmetry that a properly-informed principal might use to induce desired behavior.

Despite the pervasiveness of the phenomenon – from the media's ineffective efforts to fact-check electoral candidates, to the decisive confidence of a consultant – reflections on the difficulty of saying "I don't know" in economics at large are scant (beyond, of course, a small reputational literature which we summarize soon).[5] When experts "fake it," decision-makers may be misled into poor business decisions or bad policy choices. Still worse, trusting in the false precision of their expert reports, they may fail to see the value of investing in resources for methodical attempts to tackle hard questions, e.g. measuring the returns to advertising or evaluation of educational interventions. Or perhaps, anticipating these problems, decision-makers and the public at large learn to discount expert advice altogether. The news is uniformly bad, however: in academia, strong reputation concerns induce many to be deliberately circumspect in their claims. We hope that our positive results offer a perspective on this distinction and a fruitful direction for future work.

Structure. In Section 2 we summarize related work. In Sections 3 and 4 we present our model and equilibrium notions, respectively. Section 5 presents the non-technical intuition for our main result that difficulty validation is necessary for honesty. Sections 6 and 7 offer the technical characterization of equilibria and the relevant restrictions on the parameter space for the cases without and with policy concerns, respectively. Next, Section 8 offers comparative statics for our preferred scenario, with difficulty validation and small policy concerns. Finally, Section 9 offers a brief discussion of our results. Except where discussed explicitly in the text, all proofs are presented in Appendix B.

[5]Levitt and Dubner (2014) argue that we struggle with admitting to ourselves what we don't know, let alone to others. Motivating the problem, Steven Levitt observes "I could count on one hand the number of occasions in which someone in a company, in front of their boss, on a question that they might possibly have ever been expected to know the answer, has said 'I don't know.' Within the business world there's a general view that your job is to be an expert. And no matter how much you have to fake or how much you are making it up that you just should give an answer and hope for the best afterwards." Freakonomics Podcast – May 15, 2014. Manski (2013) offers an alternative perspective: that experts anticipate that decision-makers are "either psychologically unwilling or cognitively unable to cope with uncertainty." To be precise, Manski (2013) is concerned with the expression of uncertainty as bounds in place of point estimates, rather than the coarser but more tractable environment we study. Where these two agree is that eliciting expert uncertainty is first-order important.

2 Related Work

Since Crawford and Sobel (1982), a large literature in economics, finance, and political science has examined when informed experts can (partially) communicate their knowledge to decision-makers. In these models, a decision-maker wants to take an action which matches some state of the world, which could correspond to the effectiveness of a proposed policy or the profitability of a potential investment. Our work relates closely to a subset of this literature that focuses on professional experts – experts who have reputational concerns rather than explicit preferences over the decision-maker's action.

Prior work has shown that experts may have an incentive to bias and overstate their reports in order to convince a decision-maker that they are the "good" type (e.g., Prendergast, 1993; Prendergast and Stole, 1996; Brandenburger and Polak, 1996; Morris, 2001; Ottaviani and Sørensen, 2006a).[6] A feature of these models is that in equilibrium the behavior of the bad types may distort the incentives of the good, leading to particularly perverse outcomes (Ely and Välimäki, 2003).

Our main innovation with respect to this work is to introduce a notion of problem difficulty. Not only is the expert good or bad, but the problem may be easy or hard. In one sense this is related to the precision of expert reports in models such as Ottaviani and Sørensen (2006a), or accuracy in related and recent work on screening forecasters by Deb et al. (2018) – it creates variation in the quality of signals. However, there is an important feature that differentiates difficulty as we have formulated it – it is a property of the problem itself, not the expert or the expert's signal. This drives our results. Validating the difficulty of problems generates the key informational wedge between good uninformed experts (who know the problem difficulty) and bad experts (who do not).[7]

On the technical side, we build on results for equilibria of complete-information games in Markov strategies (Maskin and Tirole, 2001). Markov sequential equilibrium extends this notion to incomplete information games, following Bergemann and Hege (2005) and

[6] Perhaps the most common application of this argument in the literature is in models of financial forecasting (Scharfstein and Stein, 1990; Avery and Chevalier, 1999; Chevalier and Ellison, 1999; Ottaviani and Sørensen, 2006b; Hong and Kubik, 2003; Bolton et al., 2012). There is also a related literature on the value of transparency in agency relationships, where more transparency can lead to worse outcomes when bad types are no longer able to imitate the good ones (Raith and Fingleton, 2005; Prat, 2005).

[7] The closest to what we are calling difficulty in the prior literature that we are aware of is the information endowment of managers in Dye (1985) and Jung and Kwon (1988), in an altogether different setting where shareholders are uncertain as to the informational endowment of managers.

Bergemann and Hörner (2010), and we demonstrate novel implications of Markov consistency in that setting – in particular, that off-path beliefs do not depend on payoff-irrelevant information. We interpret this as a notion of *credible beliefs* which, in a game with reputational concerns, mirrors the logic of credible off-path threats (Selten, 1978).

3 The Model

Our model is motivated by the following scenario: a decision-maker (abbreviated DM, pronoun "she") is making a policy choice. There is an unknown state of the world which captures decision-relevant information – in our motivating example, whether demand is elastic or inelastic. However, the DM does not observe this state; to this end she employs an expert (pronoun "he"). Experts may be competent or incompetent. Competent experts sometimes know the state of the world, but other times the state is unknowable. Incompetent experts know nothing.

State of the World. Let the state of the world be $\omega \in \Omega \equiv \{0, 1\}$. Define p_ω to be the common knowledge probability of the *ex ante* more likely state, so $p_\omega \geq 1/2$, and without loss of generality assume this is state 1.[8] That is, $p_\omega \equiv \mathbb{P}(\omega = 1)$.

The state of the world encodes the decision-relevant information for the DM. It is unknown, and learning it is the "problem" for which she consults an expert. However, the problem is complicated for the DM by two hazards that follow directly from her lack of expertise: first, she may inadvertently hire an incompetent expert, and second, she may unwittingly ask him to solve a problem that is unsolvable.

Expert Types. The expert has a type $\theta \in \Theta \equiv \{g, b\}$, which indicates whether he is *g*ood (alternatively, "competent") or *b*ad ("incompetent"). Let $p_\theta \equiv \mathbb{P}(\theta = g)$ represent the probability that the expert is good, with $p_\theta \in (0, 1)$ and p_θ common knowledge. Experts know their type.

[8] By the symmetry of the payoffs introduced below, identical results hold if state 0 is more likely.

Problem Difficulty. The difficulty of the question is captured by another random variable $\delta \in \Delta \equiv \{e, h\}$. That is, the problem may be $e$asy (alternatively, "solvable"), or $h$ard, ("unsolvable"), where $\mathbb{P}\{\delta = e\} = p_\delta \in (0, 1)$ is the common knowledge probability of an easy problem.

The difficulty of the problem is not directly revealed to either actor at the outset. However, the expert will receive a signal, which depends on (ω, θ, δ), and good experts will be able to infer the difficulty of the problem.

Experts Signal. The expert's type and the problem difficulty determine what he learns about the state of the world. This takes the form of a signal $s \in \mathcal{S} \equiv \{s_0, s_1, s_\emptyset\}$. The expert receives a completely informative signal (i.e., $s_\omega \in \{s_0, s_1\}$) if and only if he is good *and* the problem is solvable. If not, he learns nothing about the state. Formally, let the signal be:

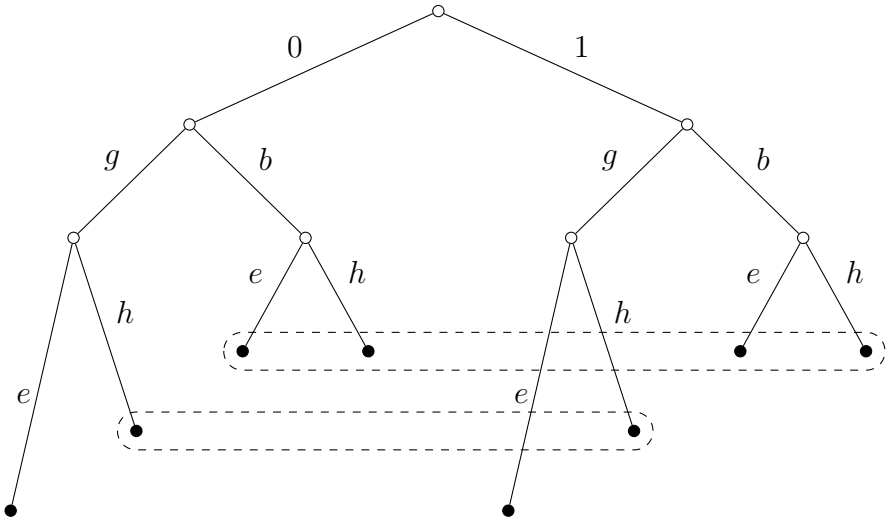
$$s = \begin{cases} s_1 & \omega = 1 \text{ , } \theta = g \text{ and } \delta = e \\ s_0 & \omega = 0 \text{ , } \theta = g \text{ and } \delta = e \\ s_\emptyset & \text{otherwise.} \end{cases} \tag{1}$$

In what follows, we will often refer to informed an uninformed experts, which is not the same as good and bad. An informed expert receives signal s_0 or s_1, and is therefore always good. However an uninformed expert receives signal $s_\emptyset$, and may be good or bad.

Sequence of Play and Validation. The game proceeds in the following sequence: first, nature plays and the state of the game (ω, θ, δ) is realized according to independent binary draws with probabilities $(p_\omega, p_\theta, p_\delta)$. Second, the expert observes his competence and signal (i.e., his information set in the second stage is $\mathcal{I}_E = (\theta, s)$), and chooses a message from a infinite message space $\mathcal{M}$. The information sets of the expert are summarized in Figure 1. There are four: first, the expert may be bad; second, the expert may be good and the problem hard; third, the expert may be good, the problem easy, and the state 0, and finally, the expert may be good, the problem easy, and the state 1.

Next the DM observes m and takes an action $a \in [0, 1]$, the *policy* choice. Her information

Figure 1: Nature's Play and Experts' Information Sets



Notes: This figure depicts Nature's moves – i.e. the choice of (ω, θ, δ) – as well as the information sets of the expert in our model. Nature moves at hollow nodes, and at solid nodes the expert choses a message. We omit the DM's policy choice as well as payoffs for simplicity.

set in this stage consists only of the expert report, i.e. $\mathcal{I}_{DM1} = (m)$.

Let $v(a, \omega)$ be the *value* of choosing policy a in state ω. We assume the policy value is given by $v(a, \omega) \equiv 1 - (a - \omega)^2$. The value of the policy is equal to 1 for making the right choice, 0 for making the worst choice, and an intermediate payoff when taking an interior action with an increasing marginal cost the further the action is from the true state. Let π_ω denote the DM's posterior belief that $\omega = 1$. Then, the expected value of taking action a is

$$1 - [\pi_\omega(1 - a)^2 + (1 - \pi_\omega)a^2], \tag{2}$$

which is maximized at $a = \pi_\omega$.

The quadratic loss formulation conveniently captures the realistic notion that when the expert does not know the state, the decision-maker makes a better policy choice (on average) when learning this rather than being misled into thinking the state is zero or one. Formally, If the problem is unsolvable, the optimal action is $a = p_\omega$, giving average payoff $1 - p_\omega(1 - p_\omega)$, which is strictly higher than the average value of the policy for any other

action.[9]

In the final stage of the game, the DM makes an inference about the expert's competence; let $\mathcal{I}_{DM2}$ represent her information set at this stage, and $\pi_\theta \equiv \mathbb{P}(\theta = g|\mathcal{I}_{DM2})$ represent the assessment of the expert. We consider several variations on $\mathcal{I}_{DM2}$. In all cases, the structure of $\mathcal{I}_{DM2}$ is assumed to be common knowledge.

In the *no validation* case, $\mathcal{I}_{DM2} = (m)$. This is meant to reflect scenarios where it is difficult or impossible to know the counterfactual outcome had the decision maker acted differently.[10]

The *state validation* case, $\mathcal{I}_{DM2} = (m, \omega)$, reflects a scenario in which the DM can check the expert's advice against the true state of the world.

This may come about *ex post* because the information is only valuable if received *ex ante*, e.g. in forecasting exercises from stock picks to elections. Alternatively, in business and policy settings, the decision-maker may be able to validate the expert's message directly, whether through experimental A/B testing of a new product feature or observational program evaluation.

In the *difficulty validation* case, $\mathcal{I}_{DM2} = (m, \delta)$, meaning that the DM learns whether the problem was hard, i.e. whether the answer could have been learned by a good expert.

There are several interpretations for this information structure. On the one hand, it could stand in for "peer review," whereby other experts evaluate the feasibility of expert's design without attempting the question themselves. Alternatively, subsequent events (such as the implementation of the policy) may reveal auxiliary information about whether the state should have been knowable, such as an extremely close election swayed by factors which should not have been *ex ante* predictable.

In the *full validation* case, $\mathcal{I}_{DM2} = (m, \omega, \delta)$, the DM learns both the state and the difficulty of the problem.[11]

[9] For example, picking $a = 0$ yields an expected value of $(1 - p_\omega)$ and $a = 1$, yielding an expected value of p_ω, both strictly less than $1 - p_\omega(1 - p_\omega)$.

[10] For instance, it is generally very difficult to estimate counterfactual outcomes to determine the value of an online advertising campaign (Gordon and Zettelmeyer, 2016).

[11] Another potentially realistic validation regime is one where the DM only learns the state if the problem is easy, i.e., has the same information as a good expert. However, the DM inference about competence when

Payoffs. The decision-maker only cares about the quality of the decision made, so

$$u_{DM} = v(a, \omega). \tag{3}$$

The expert cares about appearing competent, and potentially also about a good decision being made.[12] We parameterize his degree of *policy concerns* by $\gamma \geq 0$ and write his payoff

$$u_E = \pi_\theta + \gamma v(a, \omega). \tag{4}$$

We first consider the case where $\gamma = 0$, i.e., experts who only care about reputational concerns. We then analyze the case where $\gamma > 0$, focusing attention on the case where policy concerns are small ($\gamma \to 0$) in the main text.

To summarize, nature plays first and realizes (ω, θ, δ). Next, the expert chooses a message m given information set $\mathcal{I}_E = (\theta, s)$. Third, the DM chooses an action a given information set $\mathcal{I}_{DM1} = (m)$. Finally, the DM evaluates the competence of the expert and forms beliefs π_θ given $\mathcal{I}_{DM2}$, the manipulation of which constitutes the design problem of interest. Payoffs are $v(a, \omega)$ for the DM and $\pi_\theta + \gamma v(a, \omega)$ for the expert.

4 Equilibrium Definition and Properties

We search for *Markov sequential equilibrium* (MSE) of the model. Compared to perfect Bayesian equilibrium (PBE), this solution concept does two things. First, restricting attention to Markov strategies implies that agents making strategically equivalent decisions play the same action; in other words, their behavior cannot depend on payoff-irrelevant information. Second, it restricts off-path beliefs in a manner distinct from standard refinements. Intuitively, if agents cannot condition on payoff-irrelevant information in their actions, then consistency of beliefs implies that we cannot learn about that payoff-irrelevant information from their actions, even off-path. We introduce this solution concept to rule out unreasonable off-path beliefs necessary to sustain certain PBE; the interested reader can find a discussion of those equilibria in Appendix A.

the problem is hard does not depend on the revelation of ω, so the analysis of this case is the same as full validation.

[12] As in the career concerns literature following Holmström (1999), this payoff structure is a static representation of dynamic incentives to appear competent in order to secure future employment as an expert.

Let each node (history) be associated with information set I and an action set A. Beliefs μ map information sets into a probability distribution over their constituent nodes. Strategies σ map information sets into a probability distribution over A. Write the probability (or density) of action a at information set I as $\sigma_a(I)$. Let the function $u_I(a, \sigma)$ denote the von Neumann-Morgenstern expected utility from taking action $a \in A$ at an information set I when all subsequent play, by all players, is according to σ.

Definition 1. *A strategy σ is a **Markov strategy** if whenever, for any pair of information sets I and I' with associated action sets A and A', there exists a bijection $f : A \to A'$ such that $u_I(a, \sigma) = u_{I'}(f(a), \sigma)$, $\forall a \in A$, then $\sigma_a(I) = \sigma_{f(a)}(I')$.*[13]

The extension of equilibrium in Markov strategies to a setting with incomplete information requires some additional language. Our notation and terminology parallels the treatment of sequential equilibrium in Tadelis (2013). As consistency is to sequential equilibrium, so Markov consistency is to Markov sequential equilibrium.

Definition 2. *A profile of strategies σ and a system of beliefs μ is **Markov consistent** if there exists a sequence of non-degenerate, Markov mixed strategies $\{\sigma_k\}_{k=1}^{\infty}$ and a sequence of beliefs $\{\mu_k\}_{k=1}^{\infty}$ that are derived from Bayes' Rule, such that $\lim_{k\to\infty}(\sigma_k, \mu_k) \to (\sigma, \mu)$.*

Markov consistency has two implications. The first is a restriction to Markov strategies: players cannot condition their behavior on payoff-irrelevant private information. Anticipating the analysis that follows, it will be critical to know whether bad experts ($\theta = b$) and uninformed good experts ($\theta = g$, $\delta = h$) face a strategically equivalent choice; that is, whether knowledge of δ is payoff-relevant.

However, Markov consistency also constrains players' beliefs. In particular, it rules out off-path inferences about payoff-irrelevant information, because off-path beliefs which condition on payoff-irrelevant information can not be reached by a sequence of Markov strategies.[14] Our restriction is related to that implied by D1 and the intuitive criterion. However, where these refinements require players to make inferences about off-path play in the presence of strict differences of incentives between types, our restriction rules out inference

[13] A more general definition would only require equivalence of payoffs up to an affine transformation (Maskin and Tirole, 2001), but this is unnecessary for the exposition of our application and so we opt for a narrower, simpler definition.

[14] This is analogous to the way that sequential equilibrium restricts off-path beliefs – in that case, I cannot update, following an off-path action, on information that is not in the actor's information set; see Appendix A for further discussion.

about types in the absence of strict differences of incentives.[15] With this in hand, a notion of Markov sequential equilibrium follows directly.

Definition 3. *A profile of strategies σ, together with a set of beliefs μ, is a* ***Markov sequential equilibrium*** *if (σ^*, μ^*) is a Markov consistent perfect Bayesian equilibrium.*

In Appendix A we offer a discussion of prior usage of this solution concept, as well as an illustration of its implications for off-path beliefs designed to parallel the discussion of consistency and sequential equilibrium from Kreps and Wilson (1982).

Behavioral Motivation Markov strategies have axiomatic foundations (Harsanyi and Selten, 1988), and can be motivated by purification arguments as well as finite memory in forecasting (Maskin and Tirole, 2001; Bhaskar et al., 2013). In the complete information settings to which it is commonly applied, the Markovian restriction prevents the players from conditioning their behavior on payoff-irrelevant aspects of the (common knowledge) history.[16] The natural extension of this idea to asymmetric information games is to prevent players from conditioning on elements in their information set that are payoff-irrelevant. Or, in our setting, *types* facing strategically equivalent scenarios must play the same strategy.

Our restriction on beliefs is also related to the notion of structural consistency proposed by Kreps and Wilson (1982).[17] In that spirit, Markov consistency formalizes a notion of "credible" beliefs, analogous to the notion of credible threats in subgame perfect equilibrium. Instead of using arbitrarily punitive off-path beliefs to discipline on-path behavior, we require that off-path beliefs are credible in the sense that, *ex post*, on arriving at such an off-path node, the relevant agent could construct a convergent sequence of Markov strategies to rationalize them.

Robustness of MSE. Though the restriction to Markov strategies itself enforces a notion of robustness, there is a trivial sense in which the restriction of Markov equilibrium –

[15] For this same reason it will become apparent in Section 5 that D1 and the intuitive criterion do not have the same power to restrict off-path beliefs, see Footnote 21.

[16] Applications of Markov equilibrium have been similarly focused on the infinitely-repeated, complete information setting. See, e.g. Maskin and Tirole (1988a,b); Ericson and Pakes (1995).

[17] Kreps and Ramey (1987) demonstrated that consistency may not imply structural consistency, as conjectured by Kreps and Wilson (1982). We observe that as the Markov property is preserved by limits, Markov consistency does not introduce any further interpretive difficulty.

whether in a complete information setting or an incomplete information setting – is non-robust. That is, because it imposes symmetric strategies only when incentives are exactly symmetric, small perturbations of a model may permit much larger sets of equilibria. In the standard applications of the Markov restriction, this could be driven by future payoffs being slightly different depending on the history of play. In our setting, good and bad uninformed experts could have marginally different expected payoffs. Either way, we maintain that this is a red herring. The point of the refinement, like the symmetry condition of Nash (1950), is to hold the economist to a simple standard: that we be precise about exactly what *kind* of asymmetry in the model construction explains asymmetries in the predicted behavior. From this perspective, another interpretation of our work is that we are reflecting on exactly what informational structures introduce the asymmetry we need to obtain honesty in equilibrium.[18]

Properties of Equilibria. Since we allow for a generic message space, there will always be many equilibria to the model even with the Markov restriction. To organize the discussion, we will focus on how much information about the state and competence of the expert can be conveyed. On one extreme, we have babbling equilibria, in which all types employ the same strategy.

On the other extreme, there is never an equilibrium with full separation of types. To see why, suppose there is a message that is *only* sent by the good but uninformed types $m_{g,\emptyset}$ ("I don't know because the problem is hard") and a different message only sent by the bad uninformed types $m_{b,\emptyset}$ ("I don't know because I am incompetent"). If so, the policy choice upon observing these messages would be the same. However, for any validation regime there is some chance that a bad type can send $m_{g,\emptyset}$ and receive a strictly positive competence evaluation, and so these types have an incentive to deviate.

It will sometimes be possible to have an equilibrium where experts fully reveal their information about the *state* (but not their competence). That is, the uninformed types say "I don't know" (if not why), and the informed types report the state of the world. We call this an *honest* equilibrium.[19]

[18] We thank an anonymous referee for pointing out that one could also develop a notion of $\varepsilon-$Markov equilibrium to make this point. This is beyond the theoretical ambition of the current work, but an interesting direction for future work.

[19] Our definition of an honest equilibrium is more stringent than Sobel (1985), who only requires that good types report a message corresponding to the state. In our definition, all types must report a message which indicates their signal.

Definition 4. *Let $\pi_s(m)$ be the DM posterior belief that the expert observed signal s upon sending message m. An equilibrium is **honest** if $\pi_s(m) \in \{0,1\}\ \forall\ s \in \mathcal{S}$ and all on-path m*

It is most intuitive to formulate this equilibrium as if there were a natural language, i.e. a message m_x sent by each type observing s_x with probability 1, $x \in \{0,1,\emptyset\}$. However, more generally an equilibrium is honest if the DM always infers what signal the expert observed with certainty.

This is a particularly important class of equilibria in our model as it conveys the most information about the state possible:

Proposition 1. *The expected value of the decision in an honest equilibrium is $p_\theta p_\delta + (1 - p_\theta p_\delta)(1-(p_\omega(1-p_\omega)) \equiv \overline{v}$, which is strictly greater than the expected value of the decision in any equilibrium which is not honest.*

Proof. Unless otherwise noted, all proofs are in Appendix B. □

As in all cheap-talk games, the messages sent only convey meaning by which types send them in equilibrium. We define admitting uncertainty as sending a message which is never sent by either informed type:

Definition 5. *Let M_0 be the set of messages sent by the s_0 types and M_1 be the set of message sent by the s_1 types. Then an expert **admits uncertainty** if he sends a message $m \notin M_0 \cup M_1$*

Finally, an important class of equilibria will be one where the informed types send distinct message from each other, but the uninformed types sometimes if not always mimic these messages:

Definition 6. *A **guessing** equilibrium is one where $M_0 \cap M_1 = \emptyset$, and $Pr(m \in M_0 \cup M_1|\theta, s_\emptyset) > 0$ for at least one $\theta \in \{g,b\}$. In an **always guessing** equilibrium, $Pr(m \in M_0 \cup M_1|\theta, s_\emptyset) = 1$ for both $\theta \in \{g,b\}$.*

That is, an always guessing equilibrium is one where the informed types report their signal honestly, but the uninformed types never admit uncertainty.

Combined with proposition 1, these definitions highlight the importance of admission of uncertainty for good decision-making. In any equilibrium with guessing, the fact that the uninformed types send messages associated with informed types leads to worse policies than an honest equilibrium. This is for two reasons. First, when the expert actually is informed, their advice will be partially discounted by the fact that the DM knows some uninformed types send the informed message as well. Second, when the expert is uninformed, they will be induce the DM to take more decisive action than the expert's knowledge warrants.

5 Preliminary Observations

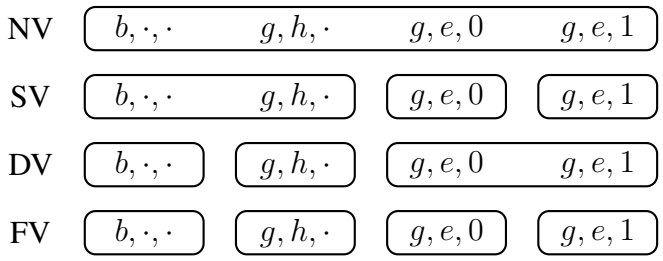
In the sections that follow we will provide a case-by case analysis of the MSE in our game for the four validations regimes, both without policy concerns (Section 6) and with (Section 7). While this is a lot to keep track of, our argument for difficulty validation has a simple structure that runs throughout the results. For the sake of exposition, here we offer the broad strokes of that argument.

Markov sequential equilibrium has two main implications: Markov strategies, implying that payoff-irrelevant information cannot affect equilibrium play, and Markov consistency, which implies in addition that off-path beliefs cannot update on payoff-irrelevant information. To see the immediate implications of these restrictions, it is helpful to construct the classes of payoff-equivalent information sets. We put information sets in the same payoff equivalence class if experts at those decision nodes are payoff-equivalent *for any DM strategy*.[20] Figure 2 illustrates for the case with no policy concerns. Each row represents an assumption on the DM's information set at the end of the game, $\mathcal{I}_{DM2}$. Each column represents one of the four information sets depicted in Figure 1.

Setting aside the parameterization $(p_\omega, p_\theta, p_\delta)$, many of our results follow directly from the structure of the payoff equivalence classes. First, in the no validation (NV) case, $\mathcal{I}_{DM2} = (m)$ and so the signal of the expert is payoff-irrelevant. There is a single payoff equivalence class comprised of all four information sets, and therefore the Markov strategies restriction implies that any MSE is a babbling equilibrium.

[20] Any two information sets can be payoff equivalent for some DM strategy: e.g., if they always pick the same policy and competence assessment for all messages.

Figure 2: Payoff Equivalence Classes With No Policy Concerns



Notes: This figure depicts equivalence classes under each validation regime for the case with no policy concerns. Each row represents a validation regime: respectively, no validation, state validation, difficulty validation, and full validation. Each column represents an expert information set, as derived in Figure 1.

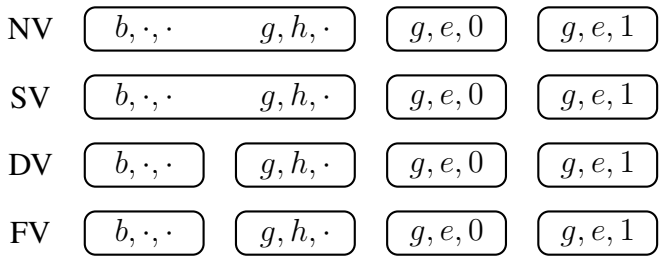
In order to sustain an honest equilibrium in Markov strategies, we need to break the payoff equivalence in a way that permits honest messages. This is exactly what state validation (SV) does, as depicted in the second row. Bad experts and uninformed good experts can pool on a message interpreted as "I don't know", and informed experts can send messages interpreted as "the state is zero" and "the state is one".

In this case, the problem is not Markov strategies but Markov consistency. Uninformed experts who deviate from saying "I don't know" risk incurring punitive beliefs if they guess incorrectly, but what can the DM credibly threaten to believe in this off-path scenario? Markov consistency bounds the severity of these beliefs because uninformed good types are payoff equivalent to bad types. In fact, the worst that the DM can threaten to believe upon observing an incorrect guess is not that the expert is bad, just that he is uninformed (i.e., in the left-most equivalence class).[21] Importantly, this is no worse than the reputational payoff associated with admitting uncertainty directly, which is what the honest equilibrium requires. Since there is a chance that the deviation is successful (if they guess the state of the world correctly), guessing is always profitable. Therefore there is *never* an honest equilibrium under state validation and no policy concerns.

For the DM to effectively threaten punitive off-path beliefs, we need to break the payoff equivalence of bad types and good but uninformed types, and this is precisely what

[21] Note here that because bad experts and uninformed good experts are strategically equivalent, D1 and the intuitive criterion do not help to restrict off-path beliefs.

Figure 3: Payoff Equivalence Classes With Policy Concerns



Notes: This figure depicts equivalence classes under each validation regime for the case with policy concerns. Each row represents a validation regime: respectively, no validation, state validation, difficulty validation, and full validation. Each column represents an expert information set, as derived in Figure 1.

difficulty validation (DV) does, depicted in the third row. Further, this effect on beliefs is complemented how DV affects the relative incentives to guess for good and bad uninformed types. With DV, admitting uncertainty is relatively palatable for good but uninformed types, who know the validation will reveal the problem is hard and give them an excuse for being uninformed. Bad types, on the other hand, do not know what the difficulty validation may reveal, and so are more tempted to (and sometimes will) guess at the state, meaning incorrect guessing is on-path and gives the expert away as incompetent.

However, difficulty validation is not enough to sustain honesty, because we also need to break the equivalence between informed experts. This we view as a more minor problem, which can be accomplished by either combining state and difficulty validation (FV), as in the fourth row, or by adding small policy concerns, which yields payoff equivalence classes represented in Figure 3.

6 No Policy Concerns

No Validation. In the case of no policy concerns and in the absence of any form of validation, the expert payoff for sending message m given their type θ and signal s is simply $\pi_\theta(m)$. This does not depend upon his private information. The restriction to Markov strategies immediately implies a strong negative result:

Proposition 2. *With no validation and no policy concerns (i.e., $\gamma = 0$), any MSE is babbling, and there is no admission of uncertainty.*

It may seem odd that even the types who know the state is zero or one can not send different messages to partially communicate this. However, if the expert does not care about the decision made, faces no checks on what he says via validation, and has no intrinsic desire to tell the truth, then his knowledge about the state is payoff-irrelevant.[22]

Small changes to the model will break the payoff equivalence of types with different knowledge about the state. In particular, any policy concerns ($\gamma > 0$) or chance that the decision-maker will learn the state will allow for separation on this dimension. However, neither of these realistic additions will change the payoff equivalence between the good and bad uninformed types who have the same knowledge about the state, which, we will show, has important implications for how much information can be transmitted in equilibrium.

State Validation. Now consider the case where the decision-maker learns the state after making their decision but before making the inference about expert competence. Write the competence assessment upon observing message m and validation ω as $\pi_\theta(m, \omega)$. The expected payoff for the expert with type and signal (θ, s) sending message m is:

$$\sum_{\omega \in \{0,1\}} \mathbb{P}(\omega|s, \theta)\pi_\theta(m, \omega)$$

–where the signal structure implies

$$\mathbb{P}(\omega = 1|\theta, s) = \begin{cases} 0 & s = s_0 \\ p_\omega & s = s_\emptyset \\ 1 & s = s_1 \end{cases}$$

–and $\mathbb{P}(\omega = 0|\theta, s) = 1 - \mathbb{P}(\omega = 1|\theta, s)$. Now the types with different information about the state are not generally payoff-equivalent since $\mathbb{P}(\omega|\theta, s)$ depends on the signal. However, good and bad uninformed experts, i.e. $(s_\emptyset, g)$ and $(s_\emptyset, b)$, are always payoff

[22] As discussed in appendix A.2, there is a PBE to the model with admission of uncertainty, though this is not sensitive to small perturbations to the model. Further, there is no honest equilibrium even without the Markov restrictions.

equivalent since $\mathbb{P}(\omega|g, s_\emptyset) = \mathbb{P}(\omega|b, s_\emptyset)$.

Given the arbitrary message space, the expert strategies can be quite complex. To search for an equilibrium as informative as possible, it is natural to start by supposing the informed types send distinct (and, for simplicity, unique) messages m_0 and m_1, respectively. We also restrict our search to messaging strategies where there is only one other message, which we label $m_\emptyset$ ("I don't know"). These restrictions are innocuous for any non-babbling equilibrium, an argument we make precise in Appendix C.[23]

So, the problem of characterizing non-babbling MSE boils down to proposing a mixed strategy over $(m_0, m_1, m_\emptyset)$ for the uninformed types, computing the relevant posterior beliefs, and checking that no type has an incentive to deviate. A mixed strategy σ is a mapping from the set of expert information sets $\{(b, s_\emptyset), (g, s_\emptyset), (g, s_0), (g, s_1)\}$ into probability weights on $\{m_\emptyset, m_0, m_1\}$, with arguments $\sigma_{\mathcal{I}_E}(m)$. In what follows we will abuse notation somewhat when describing $\mathcal{I}_E$ – for instance, where Markov strategies impose symmetric strategies between uninformed experts, we may write $\sigma_\emptyset(m_\emptyset)$ for the likelihood that an uninformed expert admits uncertainty, regardless of their competence. Moreover, in general, when an expert sends a signal m_x upon observing s_x, we will say that he sends an "honest" message.

First consider the possibility of an honest equilibrium, i.e., $\sigma_0(m_0) = \sigma_1(m_1) = \sigma_\emptyset(m_\emptyset) = 1$. In such an equilibrium, the decision-maker takes action 0 or 1 corresponding to messages m_0 and m_1, respectively, and $a = p_\omega$ when observing $m_\emptyset$. When forming a belief about the competence of the expert, state validation implies that the DM's information set is $\mathcal{I}_{DM2} = (m, \omega)$. The on-path information sets include cases where a good expert makes an accurate recommendation, $(m_0, 0)$ and $(m_1, 1)$, and cases where an uninformed expert (good or bad) says "I don't know" along with either validation result: $(m_\emptyset, 0)$, and $(m_\emptyset, 1)$. When observing $(m_0, 0)$ or $(m_1, 1)$, the DM knows the expert is good, i.e., $\pi_\theta(m_i, i) = 1$ for $i \in \{0, 1\}$. When observing $m_\emptyset$ and either validation result, the belief about the expert competence is:

$$\pi_\theta(m_\emptyset, \omega) = \frac{\mathbb{P}(\theta = g, \delta = h)}{\mathbb{P}(\theta = g, \delta = h) + \mathbb{P}(\theta = b)} = \frac{p_\theta(1 - p_\delta)}{p_\theta(1 - p_\delta) + 1 - p_\theta} \equiv \pi_\emptyset.$$

[23] This restriction is related to what Sobel (1985) calls an honest equilibrium. In Sobel (1985), "honesty" just means that the good types report a message corresponding to the state, and here we are characterizing equilibria where the good *and informed* types send a message corresponding to the state. As discussed in section 4, our definition of honesty also requires the uninformed types to always send a message corresponding to their knowledge about the state.

This expression, which recurs frequently throughout the analysis, represents the share of uninformed types who are competent (but facing a hard problem).

The informed types know that reporting their signal will yield a reputational payoff of one, and so they never have an incentive to deviate.

For the uninformed types, the expected payoff for sending $m_\emptyset$ in the honest equilibrium is $\pi_\emptyset$. Since $0 < \pi_\emptyset < p_\theta$, the expert revealing himself as uninformed leads to a lower belief about competence than the prior, but is not zero, since there are always competent but uninformed types.

Consider a deviation to m_1. When the state is in fact 1, the DM observes $(m_1, 1)$, and believes the expert to be competent with probability 1. When the state is 0, the DM observes $(m_1, 0)$, which is off-path.

However, MSE places some restriction on this belief, as it must be the limit of a sequence of beliefs consistent with a sequence of Markovian strategies. In general, the posterior belief upon observing this information set (when well-defined) is:

$$\pi_\theta(m_1, 0) = \frac{\mathbb{P}(m_1, 0, \theta = g)}{\mathbb{P}(m_1, 0)} = \frac{p_\theta p_\delta (1 - p_\omega)\sigma_0(m_1) + (1 - p_\omega)p_\theta(1 - p_\delta)\sigma_\emptyset(m_1)}{p_\theta p_\delta (1 - p_\omega)\sigma_0(m_1) + (1 - p_\omega)(1 - p_\theta p_\delta)\sigma_\emptyset(m_1)}.$$

This belief is increasing in $\sigma_0(m_1)$ and decreasing in $\sigma_\emptyset(m_1)$, and can range from $\pi_\emptyset$ (when $\sigma_0(m_1) = 0$ and $\sigma_\emptyset(m_1) > 0$) to 1 (when $\sigma_0(m_1) > 0$ and $\sigma_\emptyset(m_1) = 0$). Importantly, this lower bound results from the fact that the bad uninformed types can not send a message more often than the good uninformed types. So, upon observing an incorrect guess, MSE requires that the worst inference the DM can make is that the expert is one of the uninformed types, but can not update on which of the uninformed types is relatively more likely to guess incorrectly.[24]

Given this restriction on the off-path belief, in any honest MSE the payoff to sending m_1 must be at least:

$$p_\omega + (1 - p_\omega)\pi_\emptyset > \pi_\emptyset.$$

[24]Without the restriction to Markov beliefs, this off-path belief could be set to zero, and an honest equilibrium is sometimes possible. See appendix A.2 for further discussion of this point and why this off-path inference is fragile.

The expert can look no worse from guessing the state is one and being incorrect than they would when just admitting they are uncertain. Since there is a chance to look competent when guessing and being correct, the expert will always do so. This means there is always an incentive to deviate to m_1 (or, by an analogous argument, m_0), and hence no honest equilibrium.

A related argument shows that there is no MSE where the uninformed types *sometimes* admit uncertainty (i.e., $\sigma_\emptyset(m_\emptyset) \in (0,1)$) and sometimes guess m_0 or m_1: guessing and being correct always gives a higher competence evaluation than incorrectly guessing, which gives the same competence evaluation as sending $m_\emptyset$.

However, there is an always guessing equilibrium where the uninformed types either send only m_1 or mix between m_0 and m_1. In this equilibrium, the DM believes the expert is more likely to be competent when the state matches the message, as they either face a competent expert or an uninformed one who guessed right. Upon observing an incorrect guess, they at worst infer that the expert is uninformed if the message is off-path, and infer exactly this if the message is on-path. (Note this is exactly the somewhat-but-not-completely-punitive off-path inference used when checking for an honest MSE)

The blend of sending m_1 and m_0 depends on the probability parameters: in general, if p_ω is high the uninformed expert guesses m_1 more often if not always, as this is more likely to be what the validation reveals.

Summarizing:

Proposition 3. *With state validation and no policy concerns:*
i. In any MSE, there is no admission of uncertainty, and
ii. any non-babbling MSE is equivalent, subject to relabeling, to an MSE where both uninformed types send m_1 with probability

$$\sigma^*_\emptyset(m_1) = \begin{cases} \frac{p_\omega(1+p_\theta p_\delta)-p_\theta p_\delta}{1-p_\theta p_\delta} & \text{if} \quad p_\omega < 1/(1+p_\theta p_\delta) \\ 1 & \text{otherwise} \end{cases}$$

–and m_0 with probability $\sigma^*_\emptyset(m_0) = 1 - \sigma^*_\emptyset(m_1)$.

The always guessing MSE is more informative than babbling since the messages provide some information about the state. Further, the decision-maker learns something about the

expert's competence because upon observing a correct message the expert is more likely to be competent, and when observing an incorrect guess the decision-maker learns that the expert was uninformed (and hence less likely to be competent.)

However, since the experts never send $m_\emptyset$, there is never any *admission* of uncertainty. Put another way, the DM may learn that the expert was uninformed *ex post*, but he never says "I Don't Know." Further, the fact that the uninformed types guess dilutes the information conveyed in equilibrium.

Difficulty Validation. With difficulty validation, the informed types know that validation will reveal the problem is easy, and the good but uninformed types know that the validation will reveal the problem is hard. The bad types are unsure of what the validation will reveal.

Write the competence evaluation when the expert observes m and δ as $\pi_\theta(m, \delta)$. We can now write the expected payoff for the expert with type and signal (θ, s) sending message m as:

$$\sum_{\delta \in \{e,h\}} \mathbb{P}(\delta|s, \theta)\pi_\theta(m, \delta)$$

–where the signal structure implies

$$\mathbb{P}(\delta = e|s, \theta) = \begin{cases} 0 & s = s_\emptyset \text{ and } \theta = g \\ p_\delta & s = s_\emptyset \text{ and } \theta = b \\ 1 & s \in \{s_0, s_1\} \end{cases}$$

– and $\mathbb{P}(\delta = h|\theta, s) = 1 - \mathbb{P}(\delta = e|\theta, s)$.

In this case, the restriction to Markov strategies requires that the two informed types send the same message. Since both informed types send the same message, the DM learns nothing about the state. And since any other message is only sent by uninformed types, the DM learns nothing about the state from any message sent in an MSE.

However, with a different natural interpretation for the messages, there can be an MSE where the informed types always send a message m_e ("the problem is easy"), the good but uninformed types always send a different message m_h ("the problem is hard"), and the bad

types mix between these messages.

Proposition 4. *With no policy concerns and difficulty validation,*
i. in any MSE $a^*(m) = p_\omega$ for all on-path m, and
ii. there is an MSE where the good uninformed types always admit uncertainty.

If we care about the admission of uncertainty in and of itself, this is a positive result: difficulty validation ensures at least good uninformed types are honest about their ignorance. However, admission of uncertainty is not particularly useful if those who are informed don't reveal their information, as happens in this MSE. Put another way, one reason to desire admission of uncertainty is to avoid uninformed types diluting the value of the messages m_0 and m_1, but if there are no informative messages in the first place there is no information to dilute.

However, as we will see below, the negative aspects of the results will be fragile; combined with either state validation or any policy concerns, difficulty validation will lead to more admission of uncertainty and superior information transmission about the state.[25]

State and Difficulty (Full) Validation. While both negative in isolation, the results with just state and just difficulty validation hint at how combining the two can lead to a more positive outcome. Recalling Figure 2, with no validation, all types form one equivalence class. State validation breaks the payoff equivalence between types with different knowledge about the state, so only the good and bad uninformed types are payoff equivalent. Difficulty validation breaks the payoff equivalence with different knowledge about the difficulty of the problem, placing the informed types in the same equivalence class. Combining the two, SV and DV, no two types are payoff equivalent, which permits an honest equilibrium.

Formally, there are now four possible validation results for each message. The expected payoff to sending message m given one's type and message is:

$$\sum_{\delta\in\{e,h\}}\sum_{\omega\in\{0,1\}}\mathbb{P}(\delta|s,\theta)\mathbb{P}(\omega|s,\theta)\pi_\theta(m,\omega,\delta).$$

[25] As discussed in Appendix A, there can also be an honest equilibrium with just difficulty if we study PBE without the Markov restriction.

No pair of types share the same $\mathbb{P}(\omega|s,\theta)$ and $\mathbb{P}(\delta|s,\theta)$, so none must be payoff equivalent. As a result, all types can use distinct strategies, and off-path beliefs are unrestricted.

In an honest equilibrium, upon observing $(m_0, 0, e)$ or $(m_1, 1, e)$, the DM knows that the expert is competent. Upon observing $(m_\emptyset, \omega, e)$ the DM knows that the expert is not competent, as a competent expert would have received and sent an informative message since the problem is easy. The last on-path message/validation combination is $(m_\emptyset, \omega, h)$. Upon observing this the DM belief about the expert competence is the same as the prior, since if the problem is hard no one gets an informative message (and all send $m_\emptyset$).[26] So, the competence evaluations for the on-path messages are:

$$\pi_\theta(m_0, 0, e) = 1 \qquad \pi_\theta(m_1, 1, e) = 1$$
$$\pi_\theta(m_\emptyset, \omega, e) = 0 \qquad \pi_\theta(m_\emptyset, \omega, h) = p_\theta$$

To make honesty as easy as possible to sustain, suppose that for any off-path message ("guessing wrong"), the competence evaluation is zero. (Since no types are payoff equivalent, this belief can be rationalized as the limit of a sequence of strategies where the bad experts send m_0 and m_1 with probabilities that converge to zero more slowly than the good experts send these messages.)

The informed types get a competence evaluation of 1 for sending their honest message, so face no incentive to deviate.

A good but uninformed type knows the difficulty validation will reveal $\delta = h$, but does not know ω. Sending the honest message $m_\emptyset$ gives a competence payoff of p_θ. However, sending either m_0 or m_1 will lead to an off-path message/validation combination, and hence a payoff of zero. So, these types face no incentive to deviate.

Finally, consider the bad uninformed types, who do not know what either the state or difficulty validation will reveal. If they send $m_\emptyset$, they will be caught as uninformed if the problem was in fact easy (probability p_δ). However, if the problem is hard, the DM does not update about their competence for either state validation result. So, the expected payoff to sending $m_\emptyset$ is $(1 - p_\delta)p_\theta$.

If guessing m_1, the expert will be "caught" if either the problem is hard *or* the state is

[26] Formally, applying Bayes' rule gives $Pr(\theta = g|m_\emptyset, \omega, h) = \frac{p_\theta(1-p_\delta)}{p_\theta(1-p_\delta)+(1-p_\theta)(1-p_\delta)} = p_\theta$.

0. However, if guessing correctly, the competence evaluation will be 1. So, the expected payoff to this deviation is $p_\delta p_\omega$. Similarly, the expected payoff to guessing m_0 is $p_\delta(1 - p_\omega) < p_\delta p_\omega$.

Honesty is possible if admitting uncertainty leads to a higher competence evaluation than guessing m_1, or:

$$(1 - p_\delta)p_\theta \geq p_\delta p_\omega \implies p_\delta \leq \frac{p_\theta}{p_\theta + p_\omega}.$$

If this inequality does not hold, a fully honest MSE is not possible. However, there is always an MSE where the good but uninformed types always send $m_\emptyset$. In such an equilibrium, the bad types pick a mixed strategy over m_0, m_1, and $m_\emptyset$. Whenever the DM observes an "incorrect guess" they assign a competence evaluation of zero. So, the good uninformed types have no reason to guess since they know the problem is hard. Returning to the derivation of the honest equilibrium, the off-path beliefs in this MSE are justified, in the sense that the good types all have strict incentives to report their honest message, and the bad types are the only ones who potentially face an incentive to send m_0 or m_1 when the problem is hard or $m_\emptyset$ when the problem is easy.

Summarizing:

Proposition 5. *With no policy concerns and full validation, there is an MSE where the informed types send distinct messages and the good but uninformed types always admit uncertainty. If $p_\delta \leq \frac{p_\theta}{p_\theta + p_\omega}$, there is an honest MSE.*

This threshold has natural comparative statics. First, it is easy to maintain when p_δ is small, meaning the problem is likely to be hard. When the problem is likely to be hard, an uninformed expert is more likely to be caught guessing m_0 or m_1, and also less likely to be revealed as incompetent when sending $m_\emptyset$. Second, the threshold is easier to maintain when p_θ is high, meaning the prior is that the expert is competent. Finally, honesty is easier to sustain when p_ω is low, as this makes it more likely to be caught when guessing m_1.

7 Policy Concerns

We now consider the case where the decision-maker also cares about the policy chosen. A complete characterization of the set of equilibria for all regions of the parameter space and all validation regimes is unwieldy. In the main text we focus on what happens in the (often realistic) case where the expert primarily cares about his reputation, but also has small policy concerns. The main implication of this perturbation is that expert types with different information about the state are no longer always payoff equivalent, which allows for more communication with no validation and difficulty validation. Since the payoff equivalence of types with different information about the state is already broken by state validation and full validation, adding small policy concerns has no effect in these cases.

The appendix contains more analysis of the case where policy concerns can be larger, with an emphasis on the minimal level of policy concerns required to induce full honesty under different validation regimes.

No Validation. First consider the case with no validation and $\gamma > 0$. For a fixed DM strategy and inference about competence, the expected payoff for expert with information (θ, s) from sending message m is:

$$\pi_\theta(m) + \gamma \sum_{\omega \in \{0,1\}} \mathbb{P}(\omega|\theta, s) v(a^*(m), \omega).$$

Here, expert type enters the payoff through the $\mathbb{P}(\omega|\theta, s)$ terms. So the types observing $s_\emptyset$ are always payoff equivalent (whether good or bad), but types observing s_0 and s_1 are not. So, the restriction to Markov strategies no longer precludes an informative equilibrium, or even an honest equilibrium.

As shown in the Appendix C, it is again without loss of generality to search for equilibria with the informed types send messages m_0 and m_1, and the uninformed types send at most one other message $m_\emptyset$.

In the honest equilibrium with these messages, the payoff for an uninformed type to send

$m_\emptyset$ is

$$\pi_\emptyset + \gamma(1 - p_\omega(1 - p_\omega)). \tag{5}$$

If the expert deviates to $m \in \{m_0, m_1\}$, his payoff changes in two ways: he looks competent with probability 1 (as only competent analysts send these messages in an honest equilibrium), and the policy payoff gets worse on average. So, the payoff to choosing m_1 is:

$$1 + \gamma p_\omega. \tag{6}$$

Since $\pi_\emptyset < 1$, if γ is sufficiently small, then (6) is always greater than (5), and so the uninformed types always prefer to deviate to m_1. So, as $\gamma \to 0$, there is no honest equilibrium. A similar argument (see Appendix B) shows that there can be no equilibrium where the uninformed types ever admit uncertainty as $\gamma \to 0$.

Unlike the case with no policy concerns where any MSE is babbling, it is possible to have an always guessing MSE with small policy concerns. The equilibrium condition for such an equilibrium is that the posterior belief about expert competence is the same when sending m_0 and m_1. This is true if and only if the uninformed type send m_1 with probability p_ω.

Summarizing:

Proposition 6. *With small policy concerns ($\gamma \to 0$) and no validation:, there is a unique (subject to relabeling) always guessing equilibrium where $\sigma^*_\emptyset(m_1) \to p_\omega$ and $\sigma^*_\emptyset(m_0) \to 1 - p_\omega$.*

State Validation. Now consider the case with state validation and small policy concerns. The expected payoff for sending message m is now:

$$\sum_{\omega \in \{0,1\}} \mathbb{P}(\omega|\theta, s)(\pi_\theta(m, \omega) + \gamma v(a^*(m), \omega)).$$

So, as with state validation and $\gamma = 0$, the good and bad uninformed types are always payoff equivalent, but no other two pairs of types are. Since the payoffs for sending each message approach that of the no policy concerns case as $\gamma \to 0$, the potential equilibrium strategies are the same. There can be no admission of uncertainty because guessing m_0 or m_1 gives some chance of achieving a strictly higher competence payoff, which is worth it when γ is sufficiently small:

Proposition 7. *With small policy concerns ($\gamma \to 0$) and state validation, there is a unique always guessing equilibrium with the same strategies identified by the $\gamma = 0$ case.*

Difficulty Validation. As with the no validation case, an important difference generated by introducing any policy concerns along with difficulty validation is to break the payoff equivalence among types with different information about the state. Further, difficulty validation breaks the payoff equivalence precisely among the two types that are payoff equivalent from policy concerns alone: the good and bad uninformed types. So, in this case no two types are always payoff-equivalent.

In an honest equilibrium, the competence assessment upon observing $(m_\emptyset, h)$ is p_θ and upon observing $(m_\emptyset, e) = 0$. So, the payoff for the good but uninformed type for sending $m_\emptyset$ (who knows the validation will reveal $\delta = h$) is

$$p_\theta + \gamma(1 - p_\omega(1 - p_\omega)).$$

The bad uninformed type does not know if the validation will reveal the problem is hard, and so receives a lower expected competence evaluation and hence payoff for sending $m_\emptyset$:

$$(1 - p_\delta)p_\theta + \gamma(1 - p_\omega(1 - p_\omega)).$$

Since no types are payoff-equivalent, any off-path competence evaluations can be set to zero. In the case with only difficulty validation, these are the information sets (m_0, h) and (m_1, h), i.e., getting caught guessing about an unsolvable problem. If these are equal to zero, then a good but uninformed type knows they will look incompetent and get a worse policy upon sending either m_0 or m_1, so they have no incentive to deviate. A bad type guessing m_1 gets expected payoff:

$$p_\delta + \gamma p_\omega.$$

which is strictly higher than the payoff for sending m_0. So, the constraint for an honest

equilibrium is:

$$(1-p_\delta)p_\theta + \gamma(1-p_\omega(1-p_\omega)) \geq p_\delta + \gamma p_\omega$$
$$\gamma \geq \frac{p_\delta(1+p_\theta) - p_\theta}{(1-p_\omega)^2}.$$

As $\gamma \to 0$, this holds when:

$$p_\delta \geq \frac{p_\theta}{1+p_\theta}. \tag{7}$$

When (7) holds, then there can be a fully honest equilibrium even as $\gamma \to 0$. Importantly, *any* policy concerns, when combined with difficulty validation, can induce all uninformed experts to admit uncertainty.

If (7) does not hold, there can also be an MSE where all of the good types report honestly and the bad types play a mixed strategy over $(m_0, m_1, m_\emptyset)$. There is a more subtle incentive compatibility constraint that must be met for this equilibrium to hold: if the bad types are indifferent between sending m_0 and m_1, it can be the case that the *informed* types prefer to deviate to sending the other informed messages (i.e., the s_1 type prefers to send m_1). See the proof in Appendix B for details; in short, if the probability of a solvable problem is not too low or the probability of the state being one is not too high, then this constraint is not violated and there is an MSE where all of the good types send their honest message.[27]

Proposition 8. *As $\gamma \to 0$ with difficulty validation, there is an honest MSE if and only if $p_\delta \leq \frac{p_\theta}{1+p_\theta}$. If not, and $p_\delta \geq 2p_\omega - 1$, then there is an MSE where the good types send their*

[27]Here is an example where this constraint is violated. Suppose p_ω is close to 1, and the bad types usually send m_1, and rarely m_0. Then the tradeoff they face is that sending m_1 leads to a better policy, but a lower competence payoff when the problem is easy (when the problem is hard, the competence payoff for either guess is zero). Now consider the good expert who observes signal s_1. Compared to the bad expert, this type has a marginally stronger incentive to send m_1 (since p_ω is close to 1). However, this type *knows* that he will face a reputational loss for sending m_1 rather than m_0, while the bad type only experiences this loss with probability p_δ. So, the bad type being indifferent means the type who knows the state is 1 has a strict incentive to deviate to m_0. In general, this deviation is tough to prevent when p_δ is low and p_ω is close to 1, hence the condition in the proposition.

honest message and the bad types use the following strategy:

$$\sigma_b^*(m_\emptyset) = \begin{cases} \frac{1-p_\delta(1+p_\theta)}{1-p_\theta} & p_\delta \in \left(\frac{p_\theta}{1+p_\theta}, \frac{1}{1+p_\theta}\right) \\ 0 & p_\delta > \frac{1}{1+p_\theta}, \end{cases}, \tag{8}$$

$$\sigma_b^*(m_0) = (1-p_\omega)(1-\sigma_b^*(m_\emptyset)),$$

$$\sigma_b^*(m_1) = p_\omega(1-\sigma_b^*(m_\emptyset)).$$

Full validation. Since no pair of types is always payoff equivalent with full validation and no policy concerns, adding small policy concerns in this case has no impact on the set of MSE strategies for similar reasons as the state validation case.

8 Comparative Statics

We now ask how changing the probability parameters of the model affects the communication of uncertainty by uninformed experts, and the expected value of the DM's action in equilibrium. In principle, there are many cases to consider – four validation regimes, with no, small, or large policy concerns. Here we will focus on the three message equilibrium identified in Proposition 8 (i.e., our preferred case of difficulty validation and small policy concerns).

Better Experts Yields Better Outcomes. First consider the effect of increasing p_θ. Holding fixed expert strategies, adding more competent experts has the obvious effect of more informative messages and better decisions.

Equilibrium comparative statics on the frequency of experts admitting uncertainty are more nuanced.

To see how having more competent experts affects the unconditional probability of the expert admitting uncertainty, an important intermediate factor is whether competent types are more or less likely to send $m_\emptyset$. Given they are always honest in this equilibrium, good types will send $m_\emptyset$ with probability $1-p_\delta$. The bad types may admit uncertainty more or less frequently: in an honest equilibrium they send $m_\emptyset$ with probability 1, and for parts of

the parameter space they always guess. Plugging in their mixed strategy derived in equation 8, the bad types send $m_\emptyset$ with a probability higher than $1 - p_\delta$ if and only if $p_\delta < 1/2$.

Next, consider how *increasing* the proportion of competent experts affects the bad type strategy (recall the good types always report honestly in this equilibrium). Algebraically, the effect is obtained by differentiating (8). From this we find that the change in the probability that the bad expert sends $m_\emptyset$ is $\frac{1-2p_\delta}{(1-p_\theta)^2}$. This is positive if $p_\delta < 1/2$, negative if $p_\delta > 1/2$, and zero if $p_\delta = 1/2$. Intuitively, when p_δ is low, most of the competent types will be uninformed. Since they are honest in this equilibrium, adding more competent types tends to make admitting uncertainty more attractive for the bad types since there is a larger group of good but uninformed experts to pool with. On the other hand, when p_δ is high, competent types are usually informed, and so adding more of them makes guessing more attractive for the bad types.

Figure 4 shows how changing p_θ affects the equilibrium strategies and probability of admitting uncertainty in the top panels; and the expected value of the decision in the bottom panels. In the left panels, $p_\delta = 0.3$, and in the right panels $p_\delta = 0.7$.

Starting with the top panels, the grey line represents the bad type's probability of admitting uncertainty. As noted above, this is increasing in p_θ when p_δ is small (left panel), and eventually the equilibrium is honest.

The black line represents the *unconditional* probability of sending $m_\emptyset$. When $p_\delta < 1/2$ (left panels), for a fixed bad type strategy adding more competent experts decreases the unconditional probability of the expert admitting uncertainty. So, to the right of the dashed line (where the equilibrium is honest and hence the bad type strategy goes not change), increasing p_θ leads to less admission of uncertainty. However, to the left of the dashed line, adding more competent experts makes the bad types admit uncertainty more often. And since this part of the parameter space is when most experts are incompetent, this effect dominates and the unconditional probability of admitting uncertainty goes up.

When $p_\delta > 1/2$, (right panels), the bad types admit uncertainty less often than the good types. So, increasing p_θ leads to more admission of uncertainty for a fixed bad type strategy, as in the right part of the figure where the bad types always guess. However, the effect on the bad type strategy again works in the opposite direction for low p_θ: adding more competent experts makes the bad types guess more, and since most experts aren't competent

this leads to less admission of uncertainty.

The bottom panels show the expected value of the decision with (hypothetical) honesty using a dotted line, and the equilibrium value with a black line. These panels illustrate the more intuitive fact that as there are more competent experts, the value of the decision made is always increasing in the proportion of competent experts. The only part of the parameter space where there is a countervailing effect where adding more competent experts makes the bad types use a less informative strategy is when p_δ is high (and p_θ low), which is precisely when it is valuable to have more competent exerts since they will learn the state.

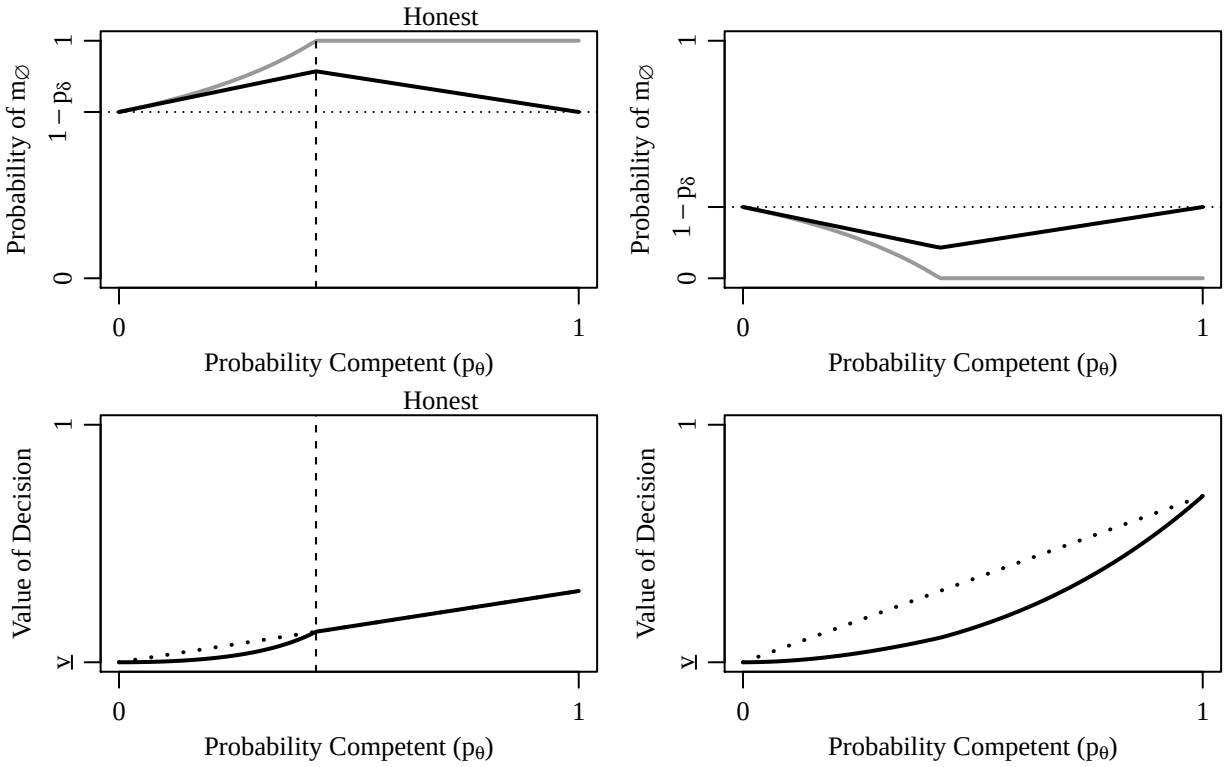
Keeping the exogenous parameters fixed, more admission of uncertainty when experts are in fact uninformed clearly leads to better decisions. However, one might expect that in environments where experts admit uncertainty more often, there is generally less information to convey, and hence worse decisions are made. Comparing the top and bottom panels highlights two scenarios where this intuition is wrong, as the frequency of admission of uncertainty and the quality of decisions move in the same direction. First, when most problems are hard and most experts are incompetent, adding more competent experts makes the modal (i.e., bad) expert more willing to admit uncertainty. Second, when most problems are easy and most experts are competent, only competent experts are willing to admit uncertainty, so adding more of them leads to hearing "I Don't Know" more often.

Formalizing these observations:

Proposition 9. *With DV and $\gamma \to 0$, in the equilibrium where the good types send the honest message:*
(i) $\mathbb{P}(m_\emptyset|\theta = b)$ is increasing in p_θ if $p_\delta > 1/2$ and decreasing p_θ if $p_\delta < 1/2$, and
(ii) The expected value of the decision is strictly increasing in p_θ.

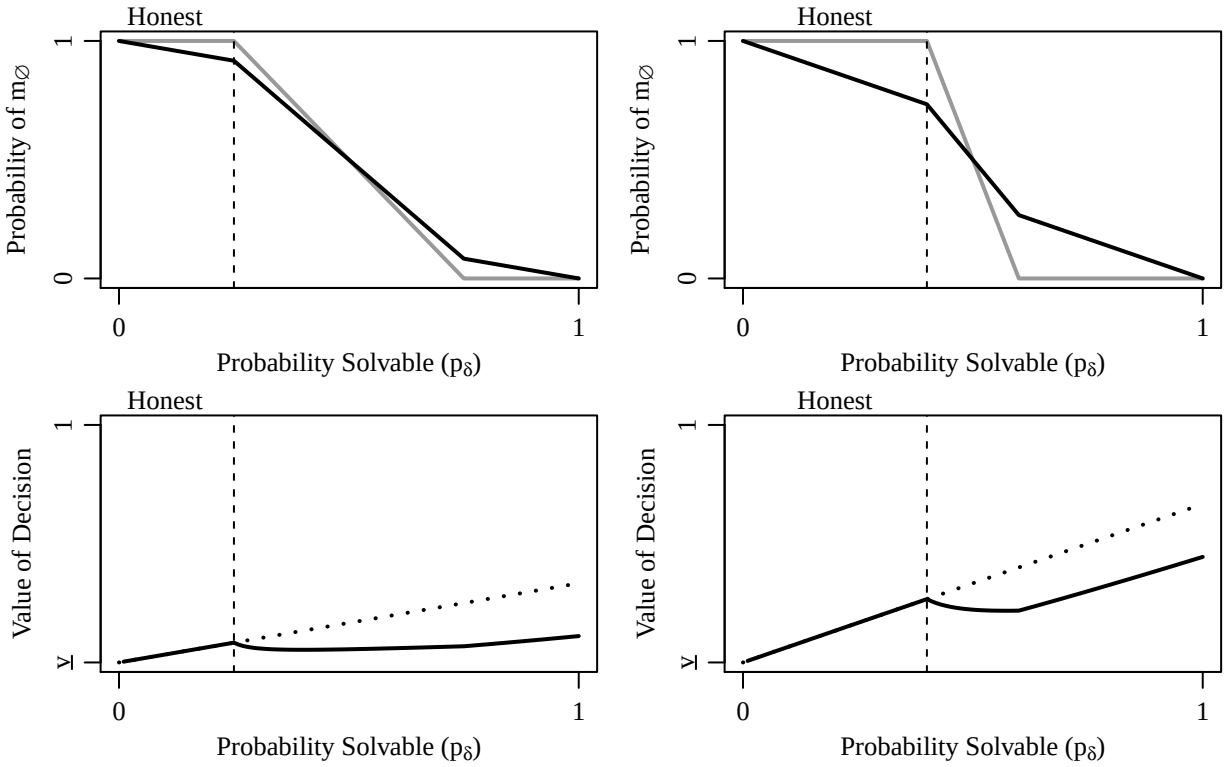
Easy Problems Can be Harder. Now consider how changing the probability that the problem is solvable affects admission of uncertainty and the value of decisions. On the admission of uncertainty, the result is straightforward: making the problem more likely to be solvable decreases the probability of sending $m_\emptyset$. This is for two reasons. First, the good experts get an informative signal more often, and hence (correctly) admit uncertainty less often. Second, the bad experts are less apt to admit uncertainty when problems are more likely to be easy. Again, this results from the fact that increasing p_δ means there is a larger pool of good experts sending informed messages and a smaller proportion sending

Figure 4: Comparative Statics in p_θ



Notes: Admission of uncertainty (top panels) and expected value of decision (bottom panels) as a function of p_θ. For the left two panels, $p_\delta = 0.3$, and for the right two panels $p_\delta = 0.7$. (p_ω does not affect these figures.) In the top panels, the black line is the unconditional probability of admitting uncertainty, and the grey line is the probability of admitting uncertainty when $\theta = b$. In the bottom panels, the dotted line is the expected value of the decision if the expert is honest, and the black line is the equilibrium expected value.

Figure 5: Comparative Statics in p_δ



Notes: Admission of uncertainty (top panels) and expected value of decision (bottom panels) as a function of p_δ. In the left panels, $p_\theta = 1/3$, and in the right panels $p_\theta = 2/3$. (p_ω does not affect these figures.) In the top panels, the black line is the unconditional probability of admitting uncertainty, and the grey line is the probability of admitting uncertainty when $\theta = b$. In the bottom panels, the dotted line is the expected value of the decision if the expert is honest, and the solid line is the equilibrium expected value.

$m_\emptyset$. The top panels of Figure 5 illustrate these claims: when p_δ is small the equilibrium is honest, and when it is large the bad types always guess. For an intermediate range, the probability of admitting uncertainty is interior and decreasing in p_δ.

While the direct and equilibrium effects move in the same direction for the admission of uncertainty, they move in opposite directions for the value of decisions. The bottom panels of Figure 5 illustrate. The expected value of the decision if the expert were to be honest (dashed line) is unsurprisingly increasing in p_δ, as the good experts are more likely to send an informative message. However, in the intermediate range where easier problems lead to more guessing, the equilibrium expected value of the decision (solid) can *decrease* as problems get easier. In this part of the parameter space, experts send

informative messages more often, but this is partly because of the fact that bad experts are guessing more. So, the decision-maker can no longer ever be confident that the state is zero or one, and this confidence decreases as problems get easier and more bad types guess. While this equilibrium effect need not always outweigh the benefits of the good experts being informed more often, for any value of p_ω and p_θ there is always *some* range of p_δ where marginal increases in p_δ to worse decisions:

Proposition 10. *With DV and $\gamma \to 0$, in the equilibrium where the good types send the honest message, there exists a $\tilde{p}_\delta \in (p_\theta/(1+p_\theta), 1/(1+p_\theta)]$ such that v^* is strictly decreasing in p_δ for $p_\delta \in (p_\theta/(1+p_\theta), \tilde{p}_\delta)$.*

9 Discussion

This paper has studied the strategic communication of uncertainty by experts with reputational concerns. Our analysis is built on two theoretical innovations: first, in our setup, the decision-maker is uncertain not only about the state of the world, but also about whether the state is *knowable*, that is, whether a qualified expert could know it. This formalizes the idea that part of being a domain expert is not merely knowing the answers to questions, but knowing how to formulate the questions themselves. The second innovation concerns the notion of "credible beliefs," which is closely tied to structural consistency of beliefs (Kreps and Wilson, 1982). Honest communication in our model is disciplined by experts' reputational concerns– off-path, they are punished by the low opinion of the decision-maker. But what can the she credibly threaten to believe? Our use of Markov sequential equilibrium restricts the decision maker to structurally consistent beliefs – that is, we do not allow the decision-maker to update on payoff-irrelevant information. We say that such beliefs are not credible because she cannot construct a candidate Markov strategy to rationalize them.

In this setting we have asked the following question: what would the decision maker want to learn, *ex post*, in order to incentivize the experts, *ex ante*, to communicate their information honestly? We found that the intuitive answer – checking experts' reports against the true state of the world – is insufficient. Even if the decision-maker catches an expert red-handed in a lie, they are constrained by the fact that good experts facing unanswerable questions are in the same conundrum as bad experts. Therefore, we show, state validation alone never induces honesty. In order to elicit honest reports from experts, it is necessary that the decision-maker also learns whether the problem is difficult. Indeed, in environments

where the expert has even very small policy concerns, difficulty validation alone may be sufficient.

What does it mean for the decision-maker to “learn the difficulty" of the problem *ex post*? On the one hand, we note that this is functionally how empirical work is evaluated in academic journals in economics. Referee reports in empirical economics typically center on questions of identification – whether a parameter is knowable in the research design – rather than the parameter value itself. However an alternative interpretation of difficulty validation concerns organizational structure and the management of experts. Should experts be allocated to product teams, managed by decision-makers who cannot evaluate their work? Or alternatively, should organizations subscribe to the “labs" model, in which experts are managed by other experts? We view our results as evidence for the latter. Perhaps it is unsurprising then, that this is precisely what firms in the tech sector – a sector opening new markets and raising new economic questions, pricing salient among them – are doing.

The literature that precedes us has shown the following to be robust: that when good experts receive imperfect signals and all experts have reputational concerns, it is difficult to incentivize honest strategic communication. We offer a simple intuition for this finding: predictive accuracy distinguishes the informed from the uninformed, not necessarily the good from the bad. If the decision-maker can learn about the problem itself they can generate informational asymmetries between good uninformed experts and bad ones, and more effectively incentivize honesty. Here we have focused on problem difficulty because we believe that decision-makers often ask unanswerable questions, but we also believe that this simple intuition may take other forms, and that further development of this idea is a fruitful area for future work.

References

Avery, C. N. and Chevalier, J. A. (1999). Herding over the career. *Economics Letters*, 63:327–333.

Bergemann, D. and Hege, U. (2005). The financing of innovation: Learning and stopping. *RAND Journal of Economics*, 36(4):719–752.

Bergemann, D. and Hörner, J. (2010). Should auctions be transparent? Cowles Foundation Discussion Paper No. 1764.

Bhaskar, V., Mailath, G. J., and Morris, S. (2013). A foundation for markov equilibria in sequential games with finite social memory. *Review of Economic Studies*, 80(3):925–948.

Bolton, P., Freixas, X., and Shapiro, J. (2012). The credit ratings game. *The Journal of Finance*, 67(1):85–111.

Brandenburger, A. and Polak, B. (1996). When managers cover their posteriors: Making the decisions the market wants to see. *The RAND Journal of Economics*, 27(3):523–541.

Chevalier, J. and Ellison, G. (1999). Career concerns of mutual fund managers. *Quarterly Journal of Economics*, 114(2):389–432.

Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica: Journal of the Econometric Society*, pages 1431–1451.

Deb, R., Pai, M., and Said, M. (2018). Evaluating strategic forecasters. forthcoming, *American Economic Review*.

Dye, R. A. (1985). Disclosure of nonproprietary information. *Journal of Accounting Research*, 23(1):123–145.

Ely, J. C. and Välimäki, J. (2003). Bad reputation. *Quarterly Journal of Economics*, 118(3):785–814.

Ericson, R. and Pakes, A. (1995). Markov-perfect industry dynamics: A framework for empirical work. *Review of Economic Studies*, 62(1):53–82.

Gordon, B. and Zettelmeyer, F. (2016). A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. Working Paper.

Harsanyi, J. C. and Selten, R. (1988). *A General Theory of Equilibrium Selection in Games*. MIT Press, Cambridge, MA.

Holmström, B. (1999). Managerial incentive problems: A dynamic perspective. *The Review of Economic Studies*, 66(1):169–182.

Hong, H. and Kubik, J. D. (2003). Analyzing the analysts: Career concerns and biased earnings forecasts. *Journal of Finance*, pages 313–351.

Jung, W.-O. and Kwon, Y. K. (1988). Disclosure when the market is unsure of the information endowment of managers. *Journal of Accounting Research*, 26(1):146–153.

Kartik, N. (2009). Strategic communication with lying costs. *Review of Economic Studies*, 76:1359–1395.

Kreps, D. M. and Ramey, G. (1987). Structural consistency, consistency, and sequential rationality. *Econometrica*, 55(6):1331–1348.

Kreps, D. M. and Wilson, R. (1982). Sequential equilibria. *Econometrica*, 50(4):863–894.

Lazear, E. P. (2005). Entrepreneurship. *Journal of Labor Economics*, 23(4):649–680.

Levitt, S. and Dubner, S. J. (2014). *Think Like a Freak*. William Morrow and Company.

Manski, C. F. (2013). *Public Policy in an Uncertain World: Analysis and Decisions*. Harvard University Press.

Maskin, E. and Tirole, J. (1988a). A theory of dynamic oligopoly, i: Overview and quantity competition with large fixed costs. *Econometrica*, 56(3):549–569.

Maskin, E. and Tirole, J. (1988b). A theory of dynamic oligopoly, ii: Price competition, kinked demand curves, and edgeworth cycles. *Econometrica*, 56(3):571–599.

Maskin, E. and Tirole, J. (2001). Markov perfect equilibriu: 1. observable actions. *Journal of Economic Theory*, 100:191–219.

Morris, S. (2001). Political correctness. *Journal of Political Economy*, 109(2):231–265.

Nash, J. F. (1950). The bargaining problem. *Econometrica*, 18(2):155–162.

Ottaviani, M. and Sørensen, P. N. (2006a). Reputational cheap talk. *RAND Journal of Economics*, 37(1).

Ottaviani, M. and Sørensen, P. N. (2006b). The strategy of professional forecasting. *Journal of Financial Economics*, 81:441–466.

Prat, A. (2005). The wrong kind of transparency. *American Economic Review*, 95(3):862–877.

Prendergast, C. (1993). A theory of "yes men". *American Economic Review*, 83(4):757–770.

Prendergast, C. and Stole, L. (1996). Impetuous youngsters and jaded old-timers: Acquiring a reputation for learning. *Journal of Political Economy*, 104(6):1105–1134.

Raith, M. and Fingleton, J. (2005). Career concerns of bargainers. *Journal of Law, Economics, and Organization*, 21(1):179–204.

Scharfstein, D. S. and Stein, J. C. (1990). Herd behavior and investment. *The American Economic Review*, pages 465–479.

Selten, R. (1978). The chain store paradox. *Theory and Decision*, 9:127–159.

Sobel, J. (1985). A theory of credibility. *The Review of Economic Studies*, pages 557–573.

Tadelis, S. (2013). *Game Theory: An Introduction*. Princeton University Press, Princeton, NJ.

A Markov Sequential Equilibrium

A.1 MSE and SE

Here we offer a brief discussion of the prior use of the Markov sequential equilibrium (MSE) solution concept as well as an illustration of its implications as a refinement on off-path beliefs.

MSE is the natural extension of Markov Perfect Equilibrium to incomplete information games. However, its usage is infrequent and sometimes informal. To our knowledge, there is no general treatment nor general guidance to the construction of the maximally coarse (Markov) partition of the action space, unlike the case of MPE (Maskin and Tirole, 2001) Bergemann and Hege (2005) and Bergemann and Hörner (2010) employ the solution concept, defining it as a perfect Bayesian equilibrium in Markovian strategies. In other words, they impose the Markov restriction only on the sequential rationality condition. This is different and rather weaker than our construction; our definition of MSE imposes the Markov assumption on both sequential rationality as well as consistency. While they do not use the Markov restriction to refine off-path beliefs, this is of no consequence for their applications.

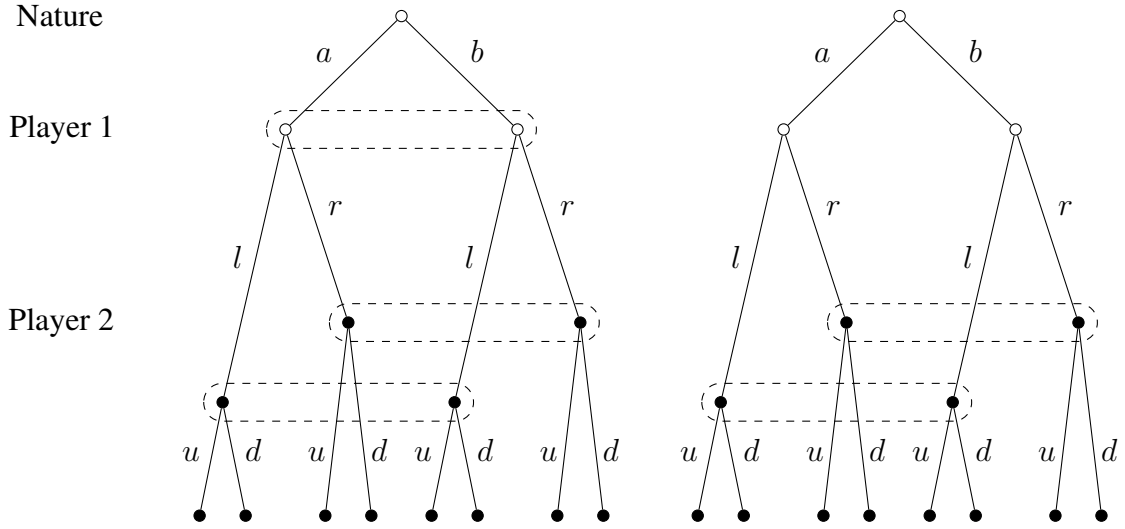
To see the relevance of MSE to off-path beliefs, consider the game illustrated in Figure A.1, which is constructed to mirror an example from Kreps and Wilson (1982).[28] First, nature choses Player 1's type, a or b. Next, Player 1 choses l or r. Finally, Player 2 choses u or d. Player 2 is never informed of Player 1's type. Whether Player 1 knows their own type is the key difference between the two games.

In the first game, the player does not know their type. Posit an equilibrium in which Player 1 always choses l. What must Player 2 believe at a node following r? If the economist is studying perfect Bayesian equilibrium (PBE), they may specify any beliefs they wish. Alternatively, if they are studying sequential equilibrium (SE), Player 2 must believe that Player 1 is of type a with probability p.

In the second game depicted, SE imposes no restriction on Player 2's off-path beliefs. However, MSE may. If $\pi_1(a, l, \cdot) = \pi_1(b, l, \cdot)$ and $\pi_1(a, r, \cdot) = \pi_1(b, r, \cdot)$ then we say

[28] See, in particular, their Figure 5 (p.873).

Figure A.1: Consistency, Markov Consistency, and Off-Path Beliefs



Notes: This figure depicts two games, which differ in whether Player 1 knows their own type. Their type, a or b, is chosen by Nature with $\mathbb{P}\{a\} = p$ and $\mathbb{P}\{b\} = 1 - p$. Player 1 chooses l or r, and Player 2 sees this and reacts with u or d. Payoffs are omitted, but can be written $\pi_i(\cdot, \cdot, \cdot)$.

that Player 1's type is *payoff irrelevant*. The restriction to Markov strategies implies that Player 1's strategy does not depend upon their type. Markov consistency implies that, further, Player 2 cannot update about payoff irrelevant information. Therefore Player 2 must believe that Player 1 is of type a with probability p.

A.2 Non-Markovian PBE

Here we briefly discuss PBE that fail the Markov consistency requirement of MSE, and argue why we believe these equilibria are less sensible

In particular, we demonstrate that the most informative equilibrium under no policy concerns and all but full validation involves more transmission of uncertainty and also information about the state. However, these equilibria are not robust to minor perturbations, such as introducing a vanishingly small cost of lying.

Example 1: Admission of Uncertainty with No Validation. Even without the Markov restriction, it is immediate that there can be no fully honest equilibrium with no validation. In such an equilibrium, the competence assessment for sending either m_0 or m_1 is 1, and the competence assessment for sending $m_\emptyset$ is $\pi_\emptyset < 1$. So the uninformed types have a strict incentive to deviate to m_0 or m_1.

However, unlike the case with the Markov restriction which leads to babbling, there is an always guessing equilibrium: If all uninformed types send m_1 with probability p_ω and m_0 otherwise, the competence assessment upon observing either message is p_θ. So no type has an incentive to deviate from the honest mesage.

Further, it is possible to get admission of uncertainty if the good and bad uninformed types play different strategies. In the extreme, suppose the good types always send their honest message, including the uninformed sending $m_\emptyset$. If the bad types were to always send m_0 or m_1, then the competence assessment upon sending $m_\emptyset$ would be 1. In this case, saying "I don't know" would lead to the highest possible competence evaluation, giving an incentive for all to admit uncertainty even if they know the state.

It is straightforward to check that if the bad types mix over message $(m_0, m_1, m_\emptyset)$ with probabilities $(p_\delta(1-p_\omega), p_\delta p_\omega, 1-p_\delta)$, then the competence assessment upon observing all messages is p_θ, and so no expert has an incentive to deviate.

A common element of these equilibria is that the competence assessment for any on-path message is equal to the prior. In fact, a messaging strategy can be part of a PBE if and only if this property holds: the competence assessments must be the same to prevent deviation, and if they are the same then by the law of iterated expectations they must equal the prior. So, there is a range of informative equilibria, but they depend on types at payoff-equivalent information sets taking different actions, a violation of Markov strategies that renders them sensitive to small perturbations of the payoffs.

Example 2: Honesty with State Validation or Difficulty Validation. Now return to the state validation case, and the conditions for an honest equilibrium. Without the Markov restriction on beliefs, it is possible to set the off-path belief upon observing an incorrect guess to 0. With this off-path belief, the incentive compatibility constraint to prevent sending m_1 becomes $\pi_\emptyset \leq p_\omega$. Since $\pi_\emptyset$ is a function of p_θ and p_δ (but not p_ω), this inequality

holds for a range of the parameter space. However, this requires beliefs that are not Markov consistent – the DM who reaches that off-path node cannot construct a Markov strategy to rationalize their beliefs. So we argue that the threat of these beliefs not credible.

Similarly, without the Markov restriction it is possible to get honesty with just difficulty validation. The binding constraint is that if any off-path message leads to a zero competence evaluation, the bad type gets a higher payoff from sending $m_\emptyset$ (as will the full validation case, $(1 - p_\delta)p_\theta$) than from sending m_1 (now p_δ). So, honesty is possible if $(1 - p_\delta)p_\theta > p_\delta$. This is a violation of Markov strategies and therefore sensitive to payoff perturbations, however in the following section we show that the same equilibrium is a MSE in the presence of small policy concerns.

The Fragility of These Examples. A standard defense of Markov strategies in repeated games is that they represent the simplest possible rational strategies (Maskin and Tirole, 2001). The similar principle applies here: rather than allowing for types with the same (effective) information to use different mixed strategies sustained by indifference, MSE focuses on the simpler case where those with the same incentives play the same strategy.

Further, as shown by Bhaskar et al. (2013) for the case of finite social memory, taking limits of vanishing, independent perturbations to the payoffs – in the spirit of Harsanyi and Selten (1988) "purification" – results in Markov strategies as well. Intuitively, suppose the expert receives a small perturbation to his payoff for sending each message which is independent of type and drawn from a continuous distribution, so he has a strict preference for sending one message over the others with probability one. Payoff-indifferent types must use the same mapping between the perturbations and messages, analogous to Markovian strategies. Further, if these perturbations put all messages on path, then all beliefs are generated by Markovian strategies.[29]

[29] A related refinement more specific to our setting is to allow for a small "lying cost" for sending a message not corresponding to the signal, which is independent of the type (Kartik, 2009).

B Proofs

Proof of Proposition 1: For convenience, we extend the definition of v so $v(a, \pi_\omega)$ represents the expected quality of policy a under the belief that the state is 1 with probability π_ω.

The DM's expected payoff from the game can be written as the sum over the (expected) payoff as a function of the expert signal:

$$\sum_{s \in \{s_0, s_1, s_\emptyset\}} \mathbb{P}(s) \sum_m \mathbb{P}(m|s) v(a^*(m), \mathbb{P}(\omega|s)). \tag{9}$$

In the honest equilibrium, when the expert observes s_0 or s_1, the DM takes an action equal to the state with probability 1, giving payoff 1. When the expert observes $s_\emptyset$, the equilibrium action is p_ω giving payoff $v(p_\omega, p_\omega) = 1 - p_\omega(1 - p_\omega)$. So, the average payoff is:

$$p_\theta p_\delta 1 + (1 - p_\theta p_\delta) p_\omega (1 - p_\omega) = \overline{v}.$$

This payoff as expressed in (9) is additively separable in the signals, and v is and globally concave in a for each s. So, for each $s \in \{s_0, s_1, s_\emptyset\}$, this component of the sum is maximized if and only if $a^*(m)$ is equal to the action taken upon observing the honest message is with probability 1. That is, it must be the case that:

$$a^*(m) = \begin{cases} 1 & m : Pr(m|s_1) > 0 \\ p_\omega & m : Pr(m|s_\emptyset) > 0 \\ 0 & m : Pr(m|s_0) > 0 \end{cases} \tag{10}$$

If the equilibrium is not honest, then there must exist a message m' such that $\mathbb{P}(s|m') < 1$ for all s. At least one of the informed types must send m' with positive probability; if not, $\mathbb{P}(s_\emptyset|m') = 1$. Suppose the type observing s_0 sends m' with positive probability. (An identical argument works if it is s_1.) To prevent $\mathbb{P}(s_0|m') = 1$ another type must send this message as well, and so in response the DM chooses an action strictly greater than 0, contradicting condition (10) and hence the expected quality of the decision in any equilibrium which is not honest is strictly less than $\overline{v}$. □

Proof of Proposition 2: For any messaging strategy, the DM must form a belief about the expert competence for any message (on- or off-path), write these $\pi_\theta(m)$. So, for any type θ, the expected utility for sending message m is just $\pi_\theta(m)$. All types are payoff-equivalent in any equilibrium, and therefore in any MSE they must use the same strategy. Since all messages are sent by both informed and uninformed types, there is no admission of uncertainty. □

Proof of Proposition 3: Part i is immediate in a babbling equilibrium: there is no admission of uncertainty since there are no messages only sent by the uninformed types. So, given propositions 12 and 13 in Appendix C, what remains to be shown is that there is no admission of uncertainty in a MSE where the s_0 types send m_0 and the s_1 types send m_1. Equivalently, in this equilibrium the uninformed types must always send m_0 or m_1.

Recall the Markov strategy restriction implies the good and bad uninformed types use the same strategy. Suppose the uninformed types send $m_\emptyset$ with positive probability. The competence assessment for sending $m_\emptyset$ is $\pi_\emptyset$. Writing the probability the uninformed types send m_1 with $\sigma_\emptyset(m_1)$, the competence assessment for sending m_1 is:

$$\begin{aligned}\mathbb{P}(\theta = g|m_1;\sigma_\emptyset(m_1)) &= \frac{p_\theta p_\delta p_\omega + p_\theta(1-p_\delta)\sigma_\emptyset(m_1)}{p_\theta p_\delta p_\omega + (p_\theta(1-p_\delta)+(1-p_\theta))\sigma_\emptyset(m_1)} \\ &\geq \frac{p_\theta p_\delta p_\omega + p_\theta(1-p_\delta)}{p_\theta p_\delta p_\omega + (p_\theta(1-p_\delta)+(1-p_\theta))} \\ &> \frac{p_\theta(1-p_\delta)}{p_\theta(1-p_\delta)+(1-p_\theta))} = \pi_\emptyset.\end{aligned}$$

Since the competence assessment for sending m_1 is strictly higher than for sending $m_\emptyset$, there can be no MSE where the uninformed types admit uncertainty, completing part i.

For part ii, first consider the condition for an equilibrium where both m_0 and m_1 are sent by the uninformed types. The uninformed types must be indifferent between guessing m_0 and m_1. This requires:

$$p_\omega \pi_\theta(m_1, \omega = 1) + (1-p_\omega)\pi_\emptyset = (1-p_\omega)\pi_\theta(m_0, \omega = 0) + p_\omega \pi_\emptyset \tag{11}$$

where the posterior beliefs upon "guessing right" are given by Bayes' rule:

$$\pi_\theta(m_1, \omega = 1) = \frac{\mathbb{P}(\theta = g, \omega = 1, m_1)}{\mathbb{P}(m_1, \omega = 1)} = \frac{p_\omega p_\theta (p_\delta + (1 - p_\delta)\sigma_\emptyset(m_1))}{p_\omega (p_\theta p_\delta + (1 - p_\theta p_\delta)\sigma_\emptyset(m_1))}$$
$$\pi_\theta(m_0, \omega = 0) = \frac{\mathbb{P}(\theta = g, \omega = 0, m_0)}{\mathbb{P}(m_0, \omega = 0)} = \frac{(1 - p_\omega) p_\theta (p_\delta + (1 - p_\delta)\sigma_\emptyset(m_0))}{(1 - p_\omega)(p_\theta p_\delta + (1 - p_\theta p_\delta)\sigma_\emptyset(m_0))}$$

Plugging these into (11) and solving for the strategies with the additional constraint that $\sigma_\emptyset(m_0) + \sigma_\emptyset(m_1) = 1$ gives:

$$\sigma_\emptyset(m_0) = \frac{1 - p_\omega(1 + p_\theta p_\delta)}{1 - p_\theta p_\delta}$$
$$\sigma_\emptyset(m_1) = \frac{p_\omega(1 + p_\theta p_\delta) - p_\theta p_\delta}{1 - p_\theta p_\delta}.$$

For this to be a valid mixed strategy, it must be the case that both of these expressions are between zero and one, which is true if and only if $p_\omega < 1/(1 + p_\theta p_\delta) \in (1/2, 1)$. So, if this inequality holds and the off-path beliefs upon observing $m_\emptyset$ are sufficiently low, there is an MSE where both messages are sent by the uninformed types. And the competence assessment for any off-path message/validation can be set to $\pi_\emptyset$, which is less than the expected competence payoff for sending either m_0 or m_1.

Now consider an equilibrium where uninformed types always send m_1. The on-path message/validation combinations are then $(m_1, \omega = 0)$, $(m_1, \omega = 1)$, and $(m_0, \omega = 0)$, with the following beliefs about the expert competence:

$$\pi_\theta(m_1, \omega = 0) = \frac{p_\theta(1 - p_\delta)}{p_\theta(1 - p_\delta) + 1 - p_\theta};$$
$$\pi_\theta(m_1, \omega = 1) = \frac{p_\theta p_\delta + p_\theta(1 - p_\delta)}{p_\theta p_\delta + p_\theta(1 - p_\delta) + (1 - p_\theta)} = p_\theta, \text{and}$$
$$\pi_\theta(m_0, \omega = 0) = 1.$$

Preventing the uninformed types from sending m_0 requires:

$$p_\omega p_\theta + (1 - p_\omega)\frac{p_\theta(1 - p_\delta)}{p_\theta(1 - p_\delta) + 1 - p_\theta} \geq p_\omega \pi_\theta(m_0, \omega = 1) + (1 - p_\omega).$$

This inequality is easiest to maintain when $\pi_\theta(m_0, \omega = 1)$ is small, and by the argument in the main text in an MSE it must be at least $\pi_\emptyset$. Setting $\pi_\theta(m_0, \omega = 1) = \pi_\emptyset$ and simplifying

gives $p_\omega \geq 1/(1+p_\theta p_\delta)$, i.e., the reverse of the inequality required for an MSE where both m_0 and m_1 are sent. Again, setting the competence assessment for an off-path message to $\pi_\emptyset$ prevents this deviation.

So, if $p_\omega \leq 1/(1+p_\theta p_\delta)$ there is an MSE where both messages are sent, and if not there is an MSE where only m_1 is sent.

Finally, it is easy to verify there is never an MSE where only m_0 is sent, as the uninformed types have an incentive to switch to m_1. $\square$

Proof of Proposition 4: Given the payoff equivalence classes, the good and informed type must used the same mixed strategy. In any MSE, the posterior belief about the state upon observing an on-path message m can be written as a weighted average of the belief about the state conditional on being in each equivalence class, weighted by the probability of being in the class:

$$\begin{aligned}\mathbb{P}(\omega=1|m) &= \mathbb{P}(\omega=1|m,\theta=g,s\in\{s_0,s_1\})\mathbb{P}(\theta=g,s\in\{s_0,s_1\}|m)\\ &\quad+\mathbb{P}(\omega=1|m,\theta=g,s=s_\emptyset)\mathbb{P}(\theta=g,s=s_\emptyset|m)\\ &\quad+\mathbb{P}(\omega=1|m,\theta=b)\mathbb{P}(\theta=b|m)\\ &= p_\omega\mathbb{P}(\theta=g,s\in\{s_0,s_1\}|m)+p_\omega\mathbb{P}(\theta=g,s=s_\emptyset|m)+p_\omega\mathbb{P}(\theta=b|m)=p_\omega.\end{aligned}$$

For each equivalence class there is no information conveyed about the state, so these conditional probabilities are all p_ω, and hence sum to this as well.

For part ii, we construct an equilibrium where the informed types always send m_e ("the problem is easy"), the good but uninformed types send m_h ("the problem is hard"), and the bad types mix over these two messages with probability $(\sigma_b(m_e),\sigma_b(m_h))$. Since m_h is never sent by the informed types, sending this message admits uncertainty.

There can be an equilibrium where both of these messages are sent by the bad types if and only if they give the same expected payoff. Writing the probability of sending m_e as $\sigma_b(m_e)$, this is possible if:

$$p_\delta\pi_\theta(m_e,e)+(1-p_\delta)\pi_\theta(m_e,h)=p_\delta\pi_\theta(m_h,e)(1-p_\delta)\pi_\theta(m_h,h),$$

– or, rearranged:

$$p_\delta \frac{p_\theta p_\delta}{p_\theta p_\delta + (1-p_\theta)\sigma_b(m_e)} = (1-p_\delta)\frac{p_\theta(1-p_\delta)}{p_\theta(1-p_\delta)+(1-p_\theta)(1-\sigma_b(m_e))}. \tag{12}$$

The left-hand side of this equation (i.e., the payoff to guessing the problem is easy) is decreasing in $\sigma_b(m_e)$, ranging from p_δ to $p_\delta \frac{p_\theta p_\delta}{p_\theta p_\delta + (1-p_\theta)}$. The right hand side is increasing in $\sigma_b(m_e)$, ranging from $(1-p_\delta)\frac{p_\theta(1-p_\delta)}{p_\theta(1-p_\delta)+(1-p_\theta)}$ to $1-p_\delta$. So, if

$$p_\delta \frac{p_\theta p_\delta}{p_\theta p_\delta + (1-p_\theta)} - (1-p_\delta) \geq 0, \tag{13}$$

then payoff to sending m_e is always higher. After multiplying through by $p_\theta p_\delta + (1-p_\theta)$, the left-hand side of (13) is quadratic in p_δ (with a positive p_δ term), and has a root at $\frac{2p_\theta - 1 + \sqrt{1+4p_\theta - 4p_\theta^2}}{4p_\theta}$ which is always on $(1/2, 1)$, and a negative root.[30] So, when p_δ is above this root, the payoff to sending m_e is always higher, and hence there is a MSE where the uninformed types always send this message.

On the other hand, if

$$(1-p_\delta)\frac{p_\theta(1-p_\delta)}{p_\theta(1-p_\delta)+(1-p_\theta)} - p_\delta \geq 0,$$

then the payoff for sending m_h is always higher, which by a similar argument holds if $p_\delta \leq \frac{2p_\theta + 1 - \sqrt{1+4p_\theta - 4p_\theta^2}}{4p_\theta}$. However, if neither of this inequalities hold, then there is a $\sigma_b(m_e) \in (0,1)$ which solves (12), and hence there is an MSE where m_e is sent with this probability and m_h with complementary probability. Summarizing, there is an MSE where the bad type sends message m_e with probability:

$$\sigma_b^*(m_e) = \begin{cases} 0 & p_\delta \leq \frac{2p_\theta + 1 - \sqrt{1+4p_\theta - 4p_\theta^2}}{4p_\theta} \\ \frac{p_\delta(p_\delta - p_\theta + 2p_\delta p_\theta - 2p_\delta^2 p_\theta)}{(1-p_\theta)(1-2p_\delta(1-p_\delta))} & p_\delta \in \left(\frac{2p_\theta + 1 - \sqrt{1+4p_\theta - 4p_\theta^2}}{4p_\theta}, \frac{2p_\theta - 1 + \sqrt{1+4p_\theta - 4p_\theta^2}}{4p_\theta} \right) \\ 1 & p_\delta \geq \frac{2p_\theta - 1 + \sqrt{1+4p_\theta - 4p_\theta^2}}{4p_\theta} \end{cases}$$

and message m_h with probability $\sigma_b^*(m_h) = 1 - \sigma_b^*(m_h)$.

[30] All of these observations follow from the fact that $1 + 4p_\theta - 4p_\theta^2 \in (1, (2p_\theta + 1)^2)$.

Proof of Proposition 5: The condition for the honest equilibrium is proven in the main text. So what remains is to show there is always an MSE where the good but uninformed type always sends $m_\emptyset$.

In such an equilibrium, message/validation combinations $(m_0, 0, e)$, $(m_1, 1, e)$ and $(m_\emptyset, 0, h)$ and $(m_\emptyset, 1, h)$ are the only ones observed when the expert is competent. So, any other message/validation combination is either on-path and only sent by the bad types, in which case the competence assessment must be 0, or is off-path and can be set to 0.

The informed type observing s_0 knows the validation will be $0, e$, and $(m, 0, e)$ leads to competence assessment zero for $m \neq m_0$. So, this type has no incentive to deviate, nor does the s_1 type by an analogous argument. The good but uninformed type knows the validation will reveal h, and the DM observing (m_i, ω, h) for $i \in \{0, 1\}$ and $\omega \in \{0, 1\}$ will lead to a competence assessment of zero. So this type faces no incentive to deviate.

What remains is showing the bad type strategy. Write the whole strategy with $\sigma_b = (\sigma_b(m_0), \sigma_b(m_1), \sigma_b(m_\emptyset))$. Explicitly deriving the conditions for all forms the (mixed) strategy can take is tedious; e.g., if p_ω is close to 1 and p_δ is close to 1, the expert always sends m_1, when p_ω is close to $1/2$ and p_δ is just below the threshold for an honest equilibrium, all three message are sent. Write the bad type's expected competence assessment for sending each message when the DM expects strategy σ (averaging over the validation result) as:

$$\begin{aligned}
\Pi_\theta(m_\emptyset, b, \sigma) &\equiv p_\delta 0 + (1 - p_\delta)\frac{p_\theta}{p_\theta + (1 - p_\theta)\sigma_b(m_\emptyset)}, \\
\Pi_\theta(m_0, b, \sigma) &\equiv p_\delta(1 - p_\omega)\frac{p_\theta}{p_\theta + (1 - p_\theta)\sigma_b(m_0)} + (1 - p_\delta)0, \text{ and} \\
\Pi_\theta(m_1, b, \sigma) &\equiv p_\delta p_\omega \frac{p_\theta}{p_\theta + (1 - p_\theta)\sigma_b(m_1)} + (1 - p_\delta)0.
\end{aligned}$$

Write the expected payoff to the bad expert choosing mixed strategy σ when the decision-maker expects mixed strategy $\hat{\sigma}_b$ as $\Pi(\sigma, \hat{\sigma}) = \sum_{i \in \{0,1,\emptyset\}} \sigma_b(m_i)\Pi_\theta(m_i; \hat{\sigma})$, which is continuous in all $\sigma_b(m_i)$, so optimizing this objective function over the (compact) unit simplex must have a solution. So, $BR(\hat{\sigma}_b) = \arg\max_\sigma \Pi(\sigma; \hat{\sigma})$ is a continuous mapping from the unit simplex to itself, which by the Kakutani fixed point theorem must have a fixed point. So, the strategy (or strategies) given by such a fixed point are a best response for the bad type when the decision-maker forms correct beliefs given this strategy. □

Proof of Proposition 6: See the proof of proposition 14 in Appendix D

Proof of Proposition 7: See the proof of proposition 15 in Appendix D

Proof of Proposition 8 See the proof of proposition 16 in Appendix D

Proof of Proposition 9: Part i is demonstrated in the main text.

For part ii, the result is immediate in the range of p_δ where p_θ does not change the bad type strategy. For the range where the bad type strategy is a function of p_θ, plugging in the strategies identified in (8) and simplifying gives the expected quality of the decision is:

$$1 - p_\omega(1 - p_\omega) + \frac{(p_\delta p_\theta)^2 p_\omega(1 - p_\omega)}{p_\delta - p_\theta(1 - 2p_\delta)}. \tag{14}$$

The derivative of (14) with respect to p_θ is:

$$\frac{p_\omega(1 - p_\omega)p_\delta^2 p_\theta(2p_\delta(1 + p_\theta) - p_\theta)}{(p_\delta - p_\theta(1 - 2p_\delta))^2}.$$

which is strictly positive if $p_\delta > \frac{p_\theta}{2(1+p_\theta)}$. Since the range of p_δ where the bad type plays a mixed strategy is $p_\delta \in (p_\theta/(1 + p_\theta), 1/(1 + p_\theta))$, this always holds. □

Proof of Proposition 10: For the range $p_\delta \in (p_\theta/(1 + p_\theta), 1/(1 + p_\theta))$, the expected quality of the decision is (14). Differentiating with respect to p_δ gives:

$$\frac{p_\omega(1 - p_\omega)p_\delta p_\theta^2(p_\delta - 2p_\theta + 2p_\delta p_\theta)}{(p_\delta - p_\theta + 2p_\delta p_\theta)^2}$$

which, evaluated at $p_\delta = p_\theta/(1 + p_\theta)$ simplifies to $-p_\omega(1 - p_\omega)$. So, the value of the decision must be locally decreasing at $p_\delta = p_\theta/(1 + p_\theta)$, and by continuity, for an open interval $p_\delta \in (p_\theta/(1 + p_\theta), \tilde{p}_\delta)$. □

C Relabeling

We prove two kinds of results in the main text. Some are existence results: that for a particular validation regime and part of the parameter space, an MSE with certain properties exists. For these results the fact that we often restrict attention to the $(m_0, m_1, m_\emptyset)$ message set poses no issues: it is sufficient to show that there is an equilibrium of this form with the claimed properties. However, propositions 2, 3, 6ii-iii, and 7, make claims that all (non-babbling) MSE have certain properties.[31] The proofs show that all equilibrium where the s_0 and s_1 types send distinct and unique messages (labelled m_0 and m_1) and there is at most one other message (labelled $m_\emptyset$) have these properties. Here we show this is WLOG in the sense that with no validation or state validation, any non-babbling equilibrium can be relabeled to an equilibrium of this form.

Consider a general messaging strategy where M is the set of messages sent with positive probability. Write the probability that the informed types observing s_0 and s_1 and $\sigma_0(m)$ and $\sigma_1(m)$. When the good and bad uninformed types are not necessarily payoff equivalent we write their strategies $\sigma_{\theta,\emptyset}(m)$. When these types are payoff equivalent and hence play the same strategy, we drop the θ: $\sigma_\emptyset(m)$. Similarly, let M_0 and M_1 be the set of messages sent by the respective informed types with strictly positive probability, and $M_{g,\emptyset}$, $M_{b,\emptyset}$, and $M_\emptyset$ the respective sets for the uninformed types, divided when appropriate.

As is standard in cheap talk games, there is always a babbling equilibrium:

Proposition 11. *There is a class of babbling equilibria where* $\sigma_0(m) = \sigma_1(m) = \sigma_{g,\emptyset}(m) = \sigma_{b,\emptyset}(m)$ *for all* $m \in M$.

Proof. If all play the same mixed strategy, then $\pi_\theta(m, \mathcal{I}_{DM2}) = p_\theta$ and $a^*(m, \mathcal{I}_{DM}) = p_\omega$ for any $m \in M$ and $\mathcal{I}_{DM}$. Setting the beliefs for any off-path message to be the same as the on-path messages, all types are indifferent between any $m \in \mathcal{M}$. □

The next result states that for all cases with either state validation or policy concerns, in any non-babbling equilibrium the informed types send no common message (note this result does *not* hold with difficulty validation; in fact the proof of proposition 4 contains a counterexample):

[31] Proposition 4 also makes a claim about all equilibria, but this is already proven in Appendix B.

Proposition 12. *With either no validation or state validation (and any level of policy concerns), any MSE where $M_0 \cap M_1 \neq \emptyset$ is babbling, i.e., $\sigma_0(m) = \sigma_1(m) = \sigma_{g,\emptyset}(m) = \sigma_{b,\emptyset}(m)$ for all $m \in M$.*

Proof. We first prove the result with state validation, and then briefly highlight the aspects of the argument that differ with no validation.

Recall that for this case the good and bad uninformed types are payoff equivalent, so we write their common message set and strategy $M_\emptyset$ and $\sigma_\emptyset(m)$ The proof proceeds in three steps.

Step 1: If $M_0 \cap M_1 \neq \emptyset$, then $M_0 = M_1$. Let $m_c \in M_0 \cap M_1$ be a message sent by both informed types. Suppose there is another message sent only by the s_0 types: $m_0 \in M_0 \backslash M_1$. For the s_0 type to be indifferent between m_0 and m_c:

$$\pi_\theta(m_c, 0) + \gamma v(a^*(m_c), 0) = \pi_\theta(m_0, 0) + \gamma v(a^*(m_0), 0).$$

For this equation to hold, it must be the case that the uninformed types send m_0 with positive probability: if not, then $\pi_\theta(m_c, 0) \leq \pi_\theta(m_0, 0) = 1$, but $v(a^*(m_c), 0) < 1 = v(a^*(m_0), 0)$, contradicting the indifference condition.

For the uninformed types to send m_0, it must also be the case that his expected payoff for sending this message, which can be written

$$p_\omega(\pi_\theta(m_0, 1) + \gamma v(a^*(m_0), 1)) + (1 - p_\omega)(\pi_\theta(m_0, 0) + v(a^*(m_0), 0))$$

– is at least his payoff for sending m_c:

$$p_\omega(\pi_\theta(m_c, 1) + \gamma v(a^*(m_c), 1)) + (1 - p_\omega)(\pi_\theta(m_c, 0) + v(a^*(m_c), 0)).$$

The second terms, which both start with $(1 - p_\omega)$, are equal by the indifference condition for s_0 types, so this requires:

$$\pi_\theta(m_0, 1) + \gamma v(a^*(m_0), 1) \geq \pi_\theta(m_c, 1) + \gamma v(a^*(m_c), 1).$$

Since m_0 is never sent by the s_1 types, $\pi_\theta(m_0, 1) = \pi_\emptyset$, while $\pi_\theta(m_c, 1) > \pi_\emptyset$. So, this

inequality requires $v(a^*(m_0), 1) > v(a^*(m_c), 1)$, which implies $a^*(m_0) > a^*(m_c)$. A necessary condition for this inequality is $\frac{\sigma_\emptyset(m_0)}{\sigma_0(m_0)} > \frac{\sigma_\emptyset(m_c)}{\sigma_0(m_c)}$, which also implies $\pi_\theta(m_c, 0) > \pi_\theta(m_0, 0)$. But if $a^*(m_0) > a^*(m_c)$ and $\pi_\theta(m_c, 0) > \pi_\theta(m_0, 0)$, the s_0 types strictly prefer to send m_c rather than m_0, a contradiction. By an identical argument, there can be no message in $M_1 \setminus M_0$, completing step 1.

Step 2: If $M_0 = M_1$, then $\sigma_0(m) = \sigma_1(m)$ for all m. If $M_0 = M_1$ is a singleton, the result is immediate. If there are multiple common messages and the informed types do not use the same mixed strategy, there must be a message m^0 such that $\sigma_0(m^0) > \sigma_1(m^0) > 0$ and another message m^1 such that $\sigma_1(m^1) > \sigma_0(m^1) > 0$. (We write the message "generally sent by type observing s_x" with a superscript to differentiate between the subscript notation referring to messages always sent by type s_x.) The action taken by the DM upon observing m^0 must be strictly less than p_ω and upon observing m^1 must be strictly greater than p_ω,[32] so $a^*(m^0) < a^*(m^1)$.

Both the s_1 and s_0 types must be indifferent between both messages, so:

$$\pi_\theta(m^0, 0) + \gamma v(a^*(m^0), 0) = \pi_\theta(m^1, 0) + \gamma v(a^*(m^1), 0)$$
$$\pi_\theta(m^0, 1) + \gamma v(a^*(m^0), 1) = \pi_\theta(m^1, 1) + \gamma v(a^*(m^1), 1)$$

Since $v(a^*(m^0), 0) > v(a^*(m^1), 0)$, for the s_0 to be indifferent it must be the case that $\pi_\theta(m^0, 0) < \pi_\theta(m^1, 0)$. Writing out this posterior belief:

$$\mathbb{P}(\theta = g|m, 0) = \frac{(1 - p_\omega)(p_\theta(p_\delta\sigma_0(m) + (1 - p_\delta)\sigma_\emptyset(m))}{(1 - p_\omega)(p_\theta p_\delta\sigma_0(m) + (1 - p_\theta p_\delta)\sigma_\emptyset(m)}.$$

Rearranging, $\pi_\theta(m^0, 0) < \pi_\theta(m^1, 0)$ if and only if $\frac{\sigma_0(m^0)}{\sigma_0(m^1)} < \frac{\sigma_\emptyset(m^0)}{\sigma_\emptyset(m^1)}$. Similarly, it must be the case that $\pi_\theta(m^1, 1) < \pi_\theta(m^0, 1)$, which implies $\frac{\sigma_1(m^0)}{\sigma_1(m^1)} > \frac{\sigma_\emptyset(m^0)}{\sigma_\emptyset(m^1)}$. Combining, $\frac{\sigma_0(m^0)}{\sigma_0(m^1)} < \frac{\sigma_1(m^0)}{\sigma_1(m^1)}$, which contradicts the definition of these messages. So, $\sigma_0(m) = \sigma_1(m)$ for all m.

Step 3: If $M_0 = M_1$ and $\sigma_0(m) = \sigma_1(m)$, then $M_\emptyset = M_0 = M_1$ and $\sigma_\emptyset(m) = \sigma_0(m) = \sigma_1(m)$. By step 2, it must be the case that $a^*(m) = p_\omega$ for all messages sent by the informed types. So, the uninformed types can't send a message not sent by the informed types: if so, the payoff would be at most $\pi_\emptyset + \gamma v(p_\omega, p_\omega)$, which is strictly less than the

[32]The action taken upon observing m can be written $\mathbb{P}(s_1|m) + p_\omega\mathbb{P}(s_\emptyset|m)$. Rearranging, this is greater than p_ω if and only if $\frac{\mathbb{P}(s_1,m)}{\mathbb{P}(s_1,m)+\mathbb{P}(s_0,m)} > p_\omega$ which holds if and only if $\sigma_1(m) > \sigma_0(m)$.

payoff for sending a message sent by the informed types. If there is only one message in M then the proof is done. If there are multiple types, all must be indifferent between each message, and by step 2 they lead to the same policy choice. So, they must also lead to the same competence assessment for each revelation of ω, which is true if and only if $\sigma_\emptyset(m) = \sigma_0(m) = \sigma_1(m)$. $\square$

Next, consider the no validation case. For step 1, define m_0 and m_1 analogously. The uninformed types must send m_0 by the same logic, and these types at least weakly prefer sending this to m_c (while the s_0 types are indifferent) requires:

$$\pi_\theta(m_0) + \gamma v(a^*(m_0), 1) \geq \pi_\theta(m_c) + \gamma v(a^*(m_c), 1).$$

This can hold only weakly to prevent the s_1 types from sending m_0 (as required by the definition). Combined with the s_0 indifference condition:

$$\pi_\theta(m_0) - \pi_\theta(m_c) = \gamma v(a^*(m_c), 1) - \gamma v(a^*(m_0), 1) = \gamma v(a^*(m_c), 0) - \gamma v(a^*(m_0), 0),$$

which requires $a^*(m_0) = a^*(m_c)$. Since the s_1 types send m_c but not m_0 this requires $\frac{\sigma_\emptyset(m_0)}{\sigma_0(m_0)} > \frac{\sigma_\emptyset(m_c)}{\sigma_0(m_c)}$, which implies $\pi_\theta(m_0) < \pi_\theta(m_c)$, contradicting the s_0 types being indifferent between both messages.

Steps 2 and 3 follow the same logic.

$\square$

Finally, we prove that any MSE where the messages sent by the s_0 and s_1 types do not overlap is equivalent to an MSE where there is only one message sent by each of these types and only one "other" message. This provides a formal statement of our claims about equilibria which are "equivalent subject to relabeling":

Proposition 13. *Let $M_U = M_\emptyset \setminus (M_0 \cup M_1)$ (i.e., the messages only sent by the uninformed types). With no validation or state validation:*
i. In any MSE where $M_0 \cap M_1 = \emptyset$, for $j \in \{0, 1, U\}$, and any $m', m'' \in M_j$, $a^*(m') = a^*(m'')$ and $\pi_\theta(m', \mathcal{I}_{DM2}) = \pi_\theta(m'', \mathcal{I}_{DM2})$
ii. Take an MSE where $|M_j| > 1$ for any $j \in \{0, 1, U\}$, and the equilibrium actions and posterior competence assessments for the messages in this set are $a^*(m_i)$ and $\pi_\theta(m_i, \mathcal{I}_{DM2})$ (which by part i are the same for all $m_i \in M_j$). Then there is another MSE where $M_j =$

$\{m\}$*, and equilibrium strategy and beliefs* a^*_{new} *and* $\pi_{\theta,new}$ *such that* $a^*(m_i) = a^*_{new}(m)$*, and* $\pi_\theta(m_i, \mathcal{I}_{DM2}) = \pi_{\theta,new}(m, \mathcal{I}_{DM2})$

Proof. For part i, first consider the message in M_U. By construction the action taken upon observing any message in this set is p_ω. And since the good and bad uninformed types are payoff equivalent and use the same strategy, the competence assessment upon observing any message in this set must be $\pi_\emptyset$.

For M_0, first note that for any $m', m'' \in M_0$, it can't be the case that the uninformed types only send one message but not the other with positive probability; if so, the message not sent by the uniformed types would give a strictly higher payoff for the s_0 types, and hence they can't send both message. So, either the uninformed types send neither m' nor m'', in which case the result is immediate, or they send both, in which case they must be indifferent between both. As shown in the proof of proposition 12, this requires that the action and competence assessment are the same for both m' and m''. An identical argument holds for M_1, completing part i.

For part ii and $M_\emptyset$, the result immediately follows from the same logic as part i.

For M_0, if the uninformed types do not send any messages in M_0, then the on-path response to any $m_0^j \in M_0$ are $a^*(m_0^j) = 0$ and $\pi_\theta(m_0^j, 0) = 1$. Keeping the rest of the equilibrium fixed, the responses in a proposed MSE where the s_0 types always send m_0 are also $a^*_{\text{new}}(m_0) = 0$ and $\pi_{\theta,\text{new}}(m_0^j, 0) = 1$. So there is an MSE where the s_0 types all send m_0 which is equivalent to the MSE where the s_0 types send multiple messages.

If the uninformed types do send the messages in M_0, then part i implies all messages must lead to the same competence evaluation, which implies for any $m_0', m_0'' \in M_0$, $\frac{\sigma_\emptyset(m_0')}{\sigma_0(m_0')} = \frac{\sigma_\emptyset(m_0'')}{\sigma_0(m_0'')} \equiv r_0$. In the new proposed equilibriuum where $M_0 = \{m_0\}$, set $\sigma_{0,\text{new}}(m_0) = 1$ and $\sigma_{\emptyset,\text{new}}(m_0) = r_0$. Since $\frac{\sigma_{\emptyset,\text{new}}(m_0)}{\sigma_{0,\text{new}}(m_0)} = \frac{\sigma_\emptyset(m_0')}{\sigma_0(m_0')}$, $a^*_{\text{new}}(m_0) = a^*(m_0')$ and $\pi_{\theta,\text{new}}(m_0', 0) = 1$, and all other aspects of the MSE are unchanged. □

D Large Policy Concerns

In the main text, we demonstrate how adding small policy concerns affects the MSE of the model under each validation regime. Not surprisingly, when policy concerns are "large", there is always an honest MSE since the expert primarily wants the DM to take the best possible action. Here we analyze how high policy concerns have to be in order to attain this honest equilibrium, and provide some results about what happens when policy concerns are not small but not large enough to induce honesty.

Since they have policy concerns, in any MSE which is not babbling, the types observing s_0 and s_1 can not send any common messages, i.e., they fully separate. Combined with a relabeling argument, for all of the analysis with policy concerns we can again restrict attention to MSE where the informed types always send m_0 and m_1, respectively, and uninformed types send at most one other message $m_\emptyset$. This is shown formally in Appendix C.

No Validation Informed types never face an incentive to deviate from the honest equilibrium: upon observing s_x for $x \in \{0, 1\}$, the DM chooses policy $a^*(s_x) = x$, and knows the expert is competent, giving the highest possible expert payoff.

Uninformed types, however, may wish to deviate. Upon observing $m_\emptyset$, the DM takes action $a = \pi_\omega = p_\omega$, which gives expected policy value $1 - p_\omega(1 - p_\omega)$, and the belief about the competence is $\pi_\emptyset$. So, for the uninformed experts of either competence type, the payoff for reporting honestly and sending signal $m_\emptyset$ is:

$$\pi_\emptyset + \gamma(1 - p_\omega(1 - p_\omega)). \tag{15}$$

If the expert deviates to $m \in \{m_0, m_1\}$, his payoff changes in two ways: he looks competent with probability 1 (as only competent analysts send these messages in an honest equilibrium), and the policy payoff gets worse on average. So, the payoff to choosing m_1 is:

$$1 + \gamma p_\omega. \tag{16}$$

As above, the payoff to deviating to m_0 is lowest, and so m_1 is the binding deviation to

check. Preventing the uninformed type from guessing m_1 requires

$$\pi_\emptyset + \gamma(1 - p_\omega(1 - p_\omega)) \geq 1 + \gamma p_\omega.$$

Rearranging, define the threshold degree of policy concerns γ^H_{NV} required to sustain honesty by

$$\begin{aligned} \gamma &\geq \frac{1 - \pi_\emptyset}{(1 - p_\omega)^2} \\ &= \frac{(1 - p_\theta)}{(1 - p_\theta p_\delta)(1 - p_\omega)^2} \\ &\equiv \gamma^H_{NV}. \end{aligned} \tag{17}$$

If $\gamma < \gamma^H_{NV}$, the uninformed types strictly prefer sending m_1 to $m_\emptyset$ *if the DM expects honesty.* Given our concern with admission of uncertainty, it is possible that there is a mixed strategy equilibrium where the uninformed types sometimes send $m_\emptyset$ and sometimes send m_0 or m_1. However, as the following result shows, when policy concerns are too small to induce full honesty, the payoff for sending m_1 is always higher than the payoff for admitting uncertainty. Moreover, since γ^H_{NV} is strictly greater than zero, when policy concerns are sufficiently small some form of validation is required to elicit any admission of uncertainty.

Proposition 14. *When $\gamma > 0$ and no validation:*
i. If $\gamma \geq \gamma^H_{NV}$, then there is an honest MSE,
ii. If $\gamma \in (0, \gamma^H_{NV})$, then all non-babbling MSE are always guessing (i.e., $\sigma^*_\emptyset(m_\emptyset) = 0$)

Proof. Part i is shown above

For part ii, it is sufficient to show that if $\gamma < \gamma^H_{NV}$, then in any proposed equilibrium where $\sigma_\emptyset(m_\emptyset) > 0$, the payoff for an expert to send m_1 is always strictly higher than the payoff to sending $m_\emptyset$. We have already shown that for this range of γ there is no honest equilibrium, i.e., if all uninformed types send $m_\emptyset$, the payoff to sending m_1 is higher than the payoff to $m_\emptyset$.

The competence evaluation upon observing m_1 as a function of the uninformed expert

mixed strategy is:

$$\pi_\theta(m_1;\sigma_\emptyset(m_1)) = \frac{\mathbb{P}(\theta = g, m_1)}{\mathbb{P}(m_1)} = \frac{p_\theta p_\omega p_\delta + p_\theta(1-p_\delta)\sigma_\emptyset(m_1)}{p_\theta p_\omega p_\delta + (p_\theta(1-p_\delta) + (1-p_\theta))\sigma_\emptyset(m_1)}$$

– and the belief about the state is:

$$\pi_\omega(m_1;\sigma_\emptyset(m_1)) = \frac{\mathbb{P}(\omega = 1, m_1)}{\mathbb{P}(m_1)} = \frac{p_\omega(p_\theta p_\delta + (1-p_\theta p_\delta)\sigma_\emptyset(m_1))}{p_\omega p_\theta p_\delta + (p_\theta(1-p_\delta) + (1-p_\theta))\sigma_\emptyset(m_1)}.$$

When observing $m_\emptyset$, the DM knows with certainty that the expert is uninformed, so $\pi_\theta(m_\emptyset) = \pi_\emptyset$ and $\pi_\omega(m_\emptyset) = p_\omega$.

Combining, the expected payoff for an uninformed type to send each message is:

$$\begin{aligned}\text{EU}(m_1; s_\emptyset, \sigma_\emptyset(m_1)) = \pi_\theta(m_1;\sigma_\emptyset(m_1)) \\ + \gamma(1 - [p_\omega(1-\pi_\omega(m_1;\sigma_\emptyset(m_1)))^2 + (1-p_\omega)\pi_\omega(m_1;\sigma_\emptyset(m_1))^2])\end{aligned}$$

– and,

$$\text{EU}(m_\emptyset) = \pi_\emptyset + \gamma(1 - p_\omega(1-p_\omega)).$$

Conveniently, $\text{EU}(m_\emptyset)$ is not a function of the mixed strategy.

If $\gamma = 0$, then $\text{EU}(m_i;\sigma_i) > \text{EU}(m_\emptyset)$ for both $i \in \{0,1\}$, because $\pi_\theta(m_i;\sigma_i) > \pi_\emptyset$. Further, by the continuity of the utility functions in γ and $\sigma_\emptyset(m_1)$, there exists a $\gamma^* > 0$ such that message m_1 will give a strictly higher payoff than $m_\emptyset$ for an open interval $(0,\gamma^*)$. The final step of the proof is to show that this γ^* is exactly γ^H_{NV}.

To show this, let $\sigma^{\text{cand}}(\gamma)$ be the candidate value of $\sigma_\emptyset(m_1)$ that solves $\text{EU}(m_1; s_\emptyset, \sigma_\emptyset(m_1)) = \text{EU}(m_\emptyset)$. Rearranging, and simplifying this equality gives:

$$\sigma^{\text{cand}}(\gamma) = -\frac{p_\omega p_\theta p_\delta}{1 - p_\theta p_\delta} + \gamma\frac{p_\omega p_\theta p_\delta (1-p_\omega)^2}{1-p_\theta}$$

which is linear in γ. When $\gamma = 0$, $\sigma^{\text{cand}}(\gamma)$ is negative, which re-demonstrates that with no policy concerns the payoff to sending m_1 is always higher than $m_\emptyset$. More generally, whenever $\sigma^{\text{cand}}(\gamma) < 0$, the payoff to sending m_1 is always higher than $m_\emptyset$ so there can be

no admission of uncertainty. Rearranging this inequality gives:

$$-\frac{p_\omega p_\theta p_\delta}{1-p_\theta p_\delta}+\gamma\frac{p_\omega p_\theta p_\delta(1-p_\omega)^2}{1-p_\theta}<0$$
$$\Leftrightarrow \gamma<\frac{1-p_\theta}{(1-p_\theta p_\delta)(1-p_\omega)^2}=\gamma^H_{NV},$$

completing part ii.

Now that we have demonstrated any equilibrium is always guessing, we can prove proposition 6. As $\gamma \to 0$, the condition for an equilibrium where the uninformed types send both m_0 and m_1 is that the competence assessments are the same. Writing these out gives:

$$\pi_\theta(m_0;\sigma_\emptyset)=\pi_\theta(m_1;\sigma_\emptyset)$$
$$\frac{p_\theta(1-p_\omega)p_\delta+p_\theta(1-p_\delta)\sigma_\emptyset(m_0)}{p_\theta(1-p_\omega)p_\delta+(p_\theta(1-p_\delta)+(1-p_\theta))\sigma_\emptyset(m_0)}=\frac{p_\theta p_\omega p_\delta+p_\theta(1-p_\delta)\sigma_\emptyset(m_1)}{p_\theta p_\omega p_\delta+(p_\theta(1-p_\delta)+(1-p_\theta))\sigma_\emptyset(m_1)}$$

which, combined with the fact that $\sigma_\emptyset(m_1)=1-\sigma_\emptyset(m_0)$ (by part ii) is true if and only if $\sigma_\emptyset(m_0)=1-p_\omega$ and $\sigma_\emptyset(m_1)=p_\omega$. There is no equilibrium where $\sigma_\emptyset(m_0)=0$; if so, $\pi_\theta(m_0;\sigma_\emptyset)=1>\pi_\theta(m_1;\sigma_\emptyset)$. Similarly, there is no equilibrium where $\sigma_\emptyset(m_1)=0$. □ □

With no validation, admission of uncertainty is now possible, though through a mechanical channel. In the extreme, when $\gamma \to \infty$, the incentives of the expert and decision-maker are fully aligned, and there is no downside to admitting uncertainty.

State Validation. Suppose there is an honest equilibrium with state validation.

As in the case with no policy concerns, upon observing message $(m_0,0)$ or $(m_1,1)$ the DM knows the expert is competent and takes an action equal to the message, and upon $(m_\emptyset,0)$ or $(m_\emptyset,1)$ takes action p_ω and knows the expert is uninformed, giving competence evaluation $\pi_\emptyset$. So, the payoff for an uninformed type to send the equilibrium message is:

$$pi_\emptyset+\gamma(1-p_\omega(1-p_\omega)). \tag{18}$$

By an identical argument to that made with no policy concerns, upon observing an off-path message, the payoff equivalence of the good and bad uninformed types implies the belief about competence in an MSE must be greater than or equal to $\pi_\emptyset$. So, the payoff to deviating to m_1 must be at least

$$p_\omega + (1 - p_\omega)\pi_\emptyset + \gamma p_\omega$$

–and the corresponding policy concerns threshold to prevent this deviation is:

$$\pi_\emptyset + \gamma(1 - p_\omega(1 - p_\omega)) \geq p_\omega + (1 - p_\omega)\pi_\emptyset + \gamma p_\omega$$

– which reduces to

$$\begin{aligned} \gamma &\geq p_\omega \gamma^H_{NV} \\ &\equiv \gamma^H_{SV} \end{aligned} \tag{19}$$

Adding state validation weakens the condition required for an honest equilibrium, particularly when p_ω is close to $1/2$. However, this threshold is always strictly positive, so for small policy concerns there can be no honesty even with state validation.

As shown in the proof of the following, if this condition is not met, then as with the no validation case there can be no admission of uncertainty. Further, since adding policy concerns does not change the classes of payoff equivalence, the case as $\gamma \to 0$ is the same as $\gamma = 0$.

Proposition 15. *With policy concerns and state validation:*
i. If $\gamma \geq \gamma^H_{SV} = p_\omega \gamma^H_{NV}$, then there is an honest MSE,
ii. If $\gamma \in (0, \gamma^H_{SV})$, then all non-babbling MSE are always guessing (i.e., $\sigma^*_\emptyset(m_\emptyset) = 0$).

Proof. Part i is demonstrated above

For part ii, our strategy mirrors the proof with no validation – that is, by way of contradiction, if the constraint for honesty is not met, then the payoff to sending m_1 is always strictly higher than $m_\emptyset$. As above, in any MSE where $\sigma_\emptyset(m_1) > 0$, the payoff for sending $m_\emptyset$ is $\pi_\emptyset + \gamma(1 - p_\omega(1 - p_\omega))$. The payoff to sending m_1 is:

$$p_\omega \pi_\theta(m_1, 1) + (1 - p_\omega)\pi_\emptyset + \gamma(1 - p_\omega(1 - \pi_\omega(m_1, \sigma_\emptyset(m_1)))^2 + (1 - p_\omega)\pi_\omega(m_1, \sigma_\emptyset(m_1))^2).$$

Next, the posterior beliefs of the decision-maker are the same as in the no validation case except:

$$\pi_\theta(m_1,1)=\frac{\mathbb{P}(\theta=g,m_1,\omega=1)}{\mathbb{P}(m_1,\omega=1)}=\frac{p_\omega p_\theta p_\delta+p_\omega p_\theta(1-p_\delta)\sigma_\emptyset(m_1)}{p_\omega p_\theta p_\delta+p_\omega(1-p_\theta p_\delta)\sigma_\emptyset(m_1)}=\frac{p_\theta p_\delta+p_\theta(1-p_\delta)\sigma_\emptyset(m_1)}{p_\theta p_\delta+(1-p_\theta p_\delta)\sigma_\emptyset(m_1)}.$$

The difference between the payoffs for sending m_1 and $m_\emptyset$ can be written:

$$p_\delta p_\theta p_\omega\frac{z(\sigma_\emptyset(m_1);\gamma)}{(1-p_\delta p_\theta)(p_\delta p_\theta(1-\sigma_\emptyset(m_1))-\sigma_\emptyset(m_1))(p_\delta p_\theta(p_\omega-\sigma_\emptyset(m_1))+\sigma_\emptyset(m_1))^2}$$

– where

$$\begin{aligned}z(\sigma_\emptyset(m_1);\gamma)=\gamma p_\delta p_\theta(-1+p_\delta p_\theta)(-1+p_\omega)^2p_\omega(p_\delta p_\theta(-1+\sigma_\emptyset(m_1))-\sigma_\emptyset(m_1))\\+(-1+p_\theta)(p_\delta p_\theta(p_\omega-\sigma_\emptyset(m_1))+\sigma_\emptyset(m_1))^2).\end{aligned}$$

So any equilibrium where both m_1 and $m_\emptyset$ are sent is characterized by $z(\sigma_\emptyset(m_1);\gamma)=0$. It is then sufficient to show that for $\gamma<\gamma^H_{SV}$, there is no $\sigma_\emptyset(m_1)\in[0,1]$ such that $z(\sigma_\emptyset(m_1);\gamma)=0$. The intuition is the same as for part ii proposition 6: the substitution effect that makes sending m_1 less appealing when other uninformed types do so is only strong when policy concerns are weak, which is precisely when sending m_1 is generally preferable to $m_\emptyset$ regardless of the uninformed type strategy.

Formally, it is easy to check that z is strictly decreasing in γ and that $z(0,\gamma^H_{SV})=0$. So, $z(0,\gamma)>0$ for $\gamma<\gamma^H_{SV}$. To show z is strictly positive for $\sigma_\emptyset(m_1)>0$, first observe that:

$$\left.\frac{\partial z}{\partial\sigma_\emptyset(m_1)}\right|_{\gamma=\gamma^H_{SV}}=(1-p_\theta)(1-p_\delta p_\theta)(p_\delta p_\theta(2-p_\omega)p_\omega+(2-2p_\delta p_\theta)\sigma_\emptyset(m_1))>0$$

– and

$$\frac{\partial^2 z}{\partial\sigma_\emptyset(m_1)\partial\gamma}=-p_\delta p_\theta(1-p_\delta p_\theta)^2(1-p_\omega)^2p_\omega<0.$$

Combined, these inequalities imply $\frac{\partial z}{\partial\sigma_\emptyset(m_1)}>0$ when $\gamma<\gamma^H_{SV}$. So, $z(\sigma_\emptyset(m_1),\gamma)>0$ for any $\sigma_\emptyset(m_1)$ when $\gamma<\gamma^H_{SV}$, completing part ii.

Now that we have proved that any equilibrium is always guessing, proposition 7 is immediate: the utilities for sending each message approach the no policy concern case as $\gamma\to0$.

□ □

Difficulty Validation. As shown in the main text, the condition for an honest equilibrium with difficulty validation and policy concerns is.

$$(1-p_\delta)p_\theta + \gamma(1-p_\omega(1-p_\omega)) \geq p_\delta + \gamma p_\omega$$
$$\gamma \geq \frac{p_\delta(1+p_\theta)-p_\theta}{(1-p_\omega)^2} \equiv \gamma^H_{DV}.$$

As discussed in the main text, γ^H_{DV} can be negative, meaning that there is an honest equilibrium even with no policy concerns. Even if not but difficulty validation also maintains a secondary advantage over state validation. Like with full validation and no policy concerns, even if it is impossible to get the *bad* uninformed types to report honestly, there is always an equilibrium where *good* uninformed types admit uncertainty. However, unlike any other case considered thus far, it is not guaranteed that the *informed* types report their honest message with positive but not too large policy concerns. See the proof in Appendix B for details; in short, if the probability of a solvable problem is not too low or the probability of the state being one is not too high, then there is an MSE where all of the good types send their honest message.[33]

Proposition 16. *With policy concerns and difficulty validation:*
i. If $\gamma \geq \gamma^H_{DV}$, then there is an honest MSE.
ii. If $\gamma \leq \gamma^H_{DV}$, then there is an MSE where the uninformed good types admit uncertainty, and if $p_\delta \geq \frac{p_\omega}{2-p_\omega}$ there is an MSE where all of the good types send their honest message.

Proof. Part i is shown above. For part ii, first note the equilibrium constructed in proposition 4 also holds with policy concerns: the policy choice upon observing both equilibrium messages is p_ω, so each type's relative payoff in this equilibrium is unaffected by the value

[33]Here is an example where this constraint is violated. Suppose p_ω is close to 1, and the bad types usually send m_1, and rarely m_0. Then the tradeoff they face is that sending m_1 leads to a better policy, but a lower competence payoff when the problem is easy (when the problem is hard, the competence payoff for either guess is zero). Now consider the good expert who observes signal s_1. Compared to the bad expert, this type has a marginally stronger incentive to send m_1 (since p_ω is close to 1). However, this type *knows* that he will face a reputational loss for sending m_1 rather than m_0, while the bad type only experiences this loss with probability p_δ. So, the bad type being indifferent means the type who knows the state is 1 has a strict incentive to deviate to m_0. In general, this deviation is tough to prevent when p_δ is low and p_ω is close to 1, hence the condition in the proposition.

of γ. Since the good uninformed types always admit uncertainty in this equilibrium, this demonstrates the first claim.

Now suppose the good types all send their honest message. By the same fixed point argument as proposition 3, the bad types must have at least one mixed strategy $(\sigma_b(m_0), \sigma_b(m_1), \sigma_b(m_\emptyset))$ which is a best response given the good types strategy and DM strategy. What remains is to show the good types have no incentive to deviate from the honest message.

The message/validation combinations (m_0, e), (m_1, e), and $(m_\emptyset, h)$ are on-path and yield competence evaluations which are all strictly greater than zero.

Message/validation combinations (m_0, h), (m_1, h), and $(m_\emptyset, e)$ are never reached with a good type. So, if the bad types send those respective messages, they are on-path and the competence assessment must be zero. If these information sets are off-path the competence assessment can be set to zero.

Since only uninformed types send $m_\emptyset$, the policy choice upon observing $m_\emptyset$ must be $a^*(m_\emptyset) = p_\omega$. The m_0 message is sent by the informed type who knows $\omega = 0$, and potentially also by uninformed bad types, so $a^*(m_0) \in [0, p_\omega)$. Similarly, $a^*(m_1) \in (p_\omega, 1]$. So $a^*(m_0) < a^*(m_\emptyset) < a^*(m_1)$.

The good and uninformed type has no incentive to deviate from sending message $m_\emptyset$ because for $m \in \{m_0, m_1\}$, $\pi_\theta(m_\emptyset, h) > \pi_\theta(m, h)$ and $v(a^*(m_\emptyset), p_\omega) > v(a^*(m), p_\omega)$.

The s_0 type has no incentive to deviate to $m_\emptyset$ since $\pi_\theta(m_0, e) > \pi_\theta(m_\emptyset, e) = 0$ and $v(a^*(m_0), 0) > v(a^*(m_\emptyset, 0))$. Similarly, the s_1 type has no incentive to deviate to $m_\emptyset$.

So, the final deviations to check are for the informed types switching to the message associated with the other state; i.e., the s_0 types sending m_1 and the s_1 types sending m_0. Preventing a deviation to m_1 requires:

$$\begin{aligned} \pi_\theta(m_0, e) + \gamma v(a^*(m_0), 0) &\geq \pi_\theta(m_1, e) + \gamma v(a^*(m_1), 0) \\ \Delta_\pi + \gamma \Delta_v(0) &\leq 0, \end{aligned} \tag{20}$$

where $\Delta_\pi \equiv \pi_\theta(m_1, e) - \pi_\theta(m_0, e)$ is the difference in competence assessments from sending m_1 versus m_0 (when the problem is easy), and $\Delta_v(p) \equiv v(a^*(m_1), p) - v(a^*(m_0), p)$ is the difference in the expected quality of the policy when sending m_1 vs m_0 for an expert

who believes $\omega = 1$ with probability p. This simplifies to:

$$\Delta_v(p) = (a^*(m_1) - a^*(m_0))(2p - a^*(m_1) - a^*(m_0)).$$

Since $a^*(m_1) > a^*(m_0)$, $\Delta_v(p)$ is strictly increasing in p, and $\Delta_v(0) < 0 < \Delta_v(1)$.

The analogous incentive compatibility constraint for the s_1 types is:

$$\Delta_\pi + \gamma\Delta_v(1) \geq 0 \tag{21}$$

If the bad types never send m_0 or m_1, then $\Delta_\pi = 0$, and (20)-(21) both hold. So, while not explicitly shown in the main text, in the honest equilibrium such a deviation is never profitable.

Now consider an equilibrium where the bad types send both m_0 and m_1, in which case they must be indifferent between both messages:

$$\begin{aligned} p_\delta\pi_\theta(m_0, e) + \gamma v(a^*(m_0), p_\omega) &= p_\delta\pi_\omega(m_1, e) + \gamma v(a^*(m_1), p_\omega) \\ p_\delta\Delta_\pi + \gamma\Delta_v(p) &= 0 \end{aligned} \tag{22}$$

Substituting this constraint into (20) and (21) and simplifying gives:

$$p_\delta\Delta_v(0) - \Delta_v(p_\omega) \leq 0 \tag{23}$$

$$p_\delta\Delta_v(1) - \Delta_v(p_\omega) \geq 0. \tag{24}$$

If $\Delta_v(p_\omega) = 0$ the constraints are both met. If $\Delta_v(p_\omega) < 0$ then the second constraint is always met, and the first constraint can be written:

$$p_\delta \geq \frac{\Delta_v(p_\omega)}{\Delta_v(0)} = \frac{a^*(m_0) + a^*(m_1) - 2p_\omega}{a^*(m_0) + a^*(m_1)} \equiv \check{p}_\delta \tag{25}$$

This constraint is hardest to meet when $\check{p}_\delta$ is large, which is true when $a^*(m_0) + a^*(m_1)$ is high. The highest value this sum can take on is $p_\omega + 1$, so $\check{p}_\delta \leq \frac{1-p_\omega}{1+p_\omega}$.

If $\Delta_v(p_\omega) > 0$, then the first constraint is always met, and the second constraint becomes:

$$p_\delta \geq \frac{\Delta_v(p_\omega)}{\Delta_v(1)} = \frac{2p_\omega - (a^*(m_0) + a^*(m_1))}{2 - (a^*(m_0) + a^*(m_1))} \equiv \hat{p}_\delta \tag{26}$$

This is hardest to meet when $a^*(m_0) + a^*(m_1)$ is small, and the smallest value it can take on is p_ω. Plugging this in, $\hat{p}_\delta \geq \frac{p_\omega}{2-p_\omega} \geq \check{p}_\delta$.

For $p_\omega \geq 1/2$, $\hat{p}_\delta \geq \check{p}_\delta$. Without placing any further restrictions on the value of $a^*(m_0) + a^*(m_1)$ – which will be straightforward on the value, this constraint ranges from $\hat{p}_\delta \in (1/3, 1)$. Still, if p_δ is sufficiently high, the informed types never have an incentive to deviate when the bad types send both m_0 and m_1.

If the bad types only send m_1 but not m_0, then the s_0 types get the highest possible payoff, so the relevant deviation to check is the s_1 types switching to m_0. The bad types sending weakly preferring m_1 implies $p_\delta \Delta_\pi + \gamma \Delta_v(p) \geq 0$, and substituting into equation 24 gives the same $p_\delta \geq \hat{p}_\delta$. Similarly, if the bad types only send m_0 but not m_1, then the relevant constraint is the s_0 types sending m_1, for which $p_\delta \geq \check{p}_\delta$ is sufficient.

Summarizing, a sufficient condition for the existence of a MSE where the good types report honestly (for any value of γ) is $p_\delta \leq p_\theta/(1+p_\theta)$ (in which case $\gamma \leq \gamma^H_{DV}$), or $p_\delta \geq \frac{p_\omega}{2-p_\omega}$. This completes part ii.

Now to prove proposition 8 we first characterize the optimal strategy for the bad types as $\gamma \to 0$, assuming the good types send their honest message. If sending $m_\emptyset$, the expert will reveal his type if $\delta = e$, but appear partially competent if $\delta = h$, giving expected payoff

$$(1 - p_\delta) \frac{p_\theta}{p_\theta + (1 - p_\theta)\sigma_b(m_\emptyset)}.$$

When sending m_0, the expert will reveal his type if $\delta = h$ (as only bad types guess when the problem is hard), but look partially competent if $\delta = e$:

$$p_\delta \frac{p_\theta(1 - p_\omega)}{p_\theta(1 - p_\omega) + (1 - p_\theta)\sigma_b(m_0)}.$$

and when sending m_1 the expect payoff is:

$$p_\delta \frac{p_\theta p_\omega}{p_\theta p_\omega + (1 - p_\theta)\sigma_b(m_1)}.$$

setting these three equal subject to $\sigma_b(m_0) + \sigma_b(m_1) + \sigma_b(m_\emptyset) = 1$ gives:

$$\begin{aligned}\sigma_b(m_\emptyset) &= \frac{1 - p_\delta(1+p_\theta)}{1-p_\theta};\\ \sigma_b(m_0) &= \frac{(1-p_\omega)(p_\delta - p_\theta(1-p_\delta))}{1-p_\theta}\\ \sigma_b(m_1) &= \frac{p_\omega(p_\delta - p_\theta(1-p_\delta))}{1-p_\theta}.\end{aligned}$$

These are all interior if and only if:

$$0 < \frac{1 - p_\delta(1+p_\theta)}{1-p_\theta} < 1 \Longrightarrow \frac{p_\theta}{1+p_\theta} < p_\delta < \frac{1}{1+p_\theta}.$$

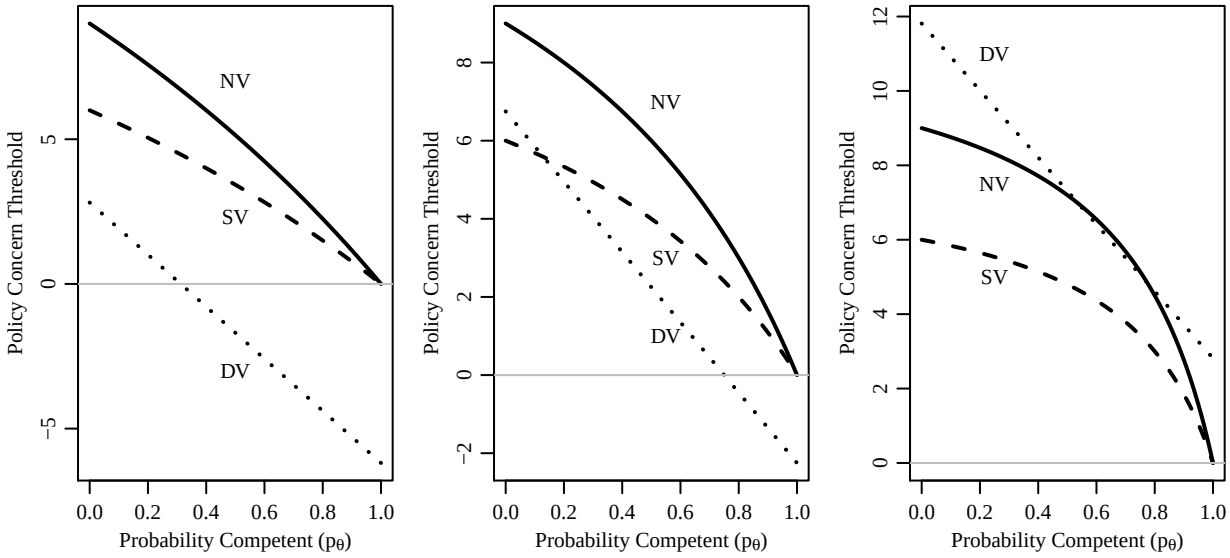
If $p_\delta \leq \frac{p_\theta}{1+p_\theta}$, then there is no fully mixed strategy for the bad expert in equilibrium because they would always prefer to send $m_\emptyset$; and recall this is exactly the condition for an honest equilibrium with no validation. If $p_\delta \geq \frac{1}{1+p_\theta}$, then the bad type always guesses. Setting the payoff for a bad type sending m_0 and m_1 equal along with $\sigma_b(m_0) + \sigma_b(m_1) = 1$ gives the strategies in the statement of the proposition.

The final step is to ensure the informed types do not send the message associated with the other state. Recall the IC constraints depend on $a^*(m_0) + a^*(m_1)$, which we can now restrict to a narrower range given the bad type strategy:

$$\begin{aligned}a^*(m_0) + a^*(m_1) &= \frac{(1-p_\theta)p_\omega(1-p_\omega)(1-\sigma_b(m_\emptyset))}{p_\delta p_\theta(1-p_\omega + (1-p_\theta)(1-p_\omega)(1-\sigma_b(m_\emptyset)))}\\ &\quad + \frac{p_\delta p_\theta p_\omega + (1-p_\theta)p_\omega p_\omega(1-\sigma_b)(m_\emptyset)}{p_\delta p_\theta p_\omega + (1-p_\theta)p_\omega(1-\sigma_b)(m_\emptyset)}\\ &= \frac{p_\delta p_\theta + (1-\sigma_b(m_\emptyset))(1-p_\theta)2p_\omega}{p_\delta p_\theta + (1-\sigma_b(m_\emptyset))(1-p_\theta)}.\end{aligned}$$

This can be interpreted as weighted average of 1 (with weight $p_\delta p_\theta$) and $2p_\omega > 1$ (with weight $(1 - \sigma_b(m_\emptyset)(1-p_\theta))$, and so must lie on $[1, 2p_\omega]$. So, (26) is always the binding constraint, and is hardest to satisfy when $a^*(m_0)+a^*(m_1) \to 1$, in which case the constraint becomes $\hat{p}_\delta = 2p_\omega - 1$. So, $p_\delta \geq 2p_\omega - 1$ is a sufficient condition for the informed types to never deviate. For any $p_\delta > 0$, this holds for p_ω sufficiently close to $1/2$, completing part proposition 8. □

Figure D.2: Comparative Statics of Honesty Threshold



Notes: Comparison of threshold in policy concerns for full honesty under different validation regimes as a function of p_θ. The panels vary in the likelihood the problem is solvable, which is 0.25 in the left panel, 0.5 in the middle panel, and 0.75 in the right panel.

Comparative Statics: Difficulty Validation Can be the Wrong Kind of Transparency. As long as policy concerns are strictly positive but small, difficulty validation is more effective at eliciting honesty than state validation.

For larger policy concerns the comparison becomes less straightforward. Figure D.2 shows the policy concern threshold for honesty under no validation (solid line), state validation (dashed line), and difficulty validation (dotted line) as a function of the prior on the expert competence, when the problem is usually hard ($p_\delta = 0.25$, left panel), equally likely to be easy or hard ($p_\delta = 0.5$, middle panel) and usually easy ($p_\delta = 0.75$, right panel). In all panels $p_\omega = 0.67$; changing this parameter does not affect the conclusions that follow.[34] For intuition, difficulty validation makes it hard to compensate bad experts for saying "I don't know," as there are fewer good experts who don't know. For very easy problems difficulty validation can be *worse* than no validation. This mirrors the result in Prat (2005), where transparency can eliminate incentives for bad types to pool with good types by exerting more effort.

[34] In general, honesty is easier to sustain under all validation regimes when p_ω is lower, with state validation being particularly sensitive to this change.

This figure illustrates several key conclusions from the model. First, in all cases, the policy concern threshold required is decreasing in p_θ, which means it is easier to sustain honesty when the prior is that the expert is competent. This is because when most experts are competent in general, most uninformed experts are competent as well, and so there is less of a penalty for admitting uncertainty. Second, the threshold with state validation is always lower than the threshold with no validation, though these are always strictly positive as long as $p_\theta < 1$. Further, for most of the parameter space these thresholds are above two, indicating the expert must care twice as much about policy than about perceptions of his competence to elicit honesty. On the other hand, in the right and middle panels there are regions where the threshold with difficulty validation is below zero, indicating no policy concerns are necessary to induce admission of uncertainty (in fact, the expert could want the decision-maker to make a *bad* decision and still admit uncertainty).

Finally, consider how the relationship between the thresholds changes as the problem becomes easier. When problems are likely to be hard (left panel), difficulty validation is the best for eliciting honesty at all values of p_θ. In the middle panel, difficulty validation is always better than no validation, but state validation is best for low values of p_θ. When the problem is very likely to be easy, difficulty validation is always worse than state validation and is even worse than even no validation other than for a narrow range of p_θ.

However, even in this case difficulty validation still can elicit honesty from good but uninformed experts when policy concerns are not high enough, while there is no admission of uncertainty at all when policy concerns are not high enough with no validation and state validation.