Good cop-bad cop: delegating interrogations

Alessandro Ispano Péter Vida

July 2022

Abstract

A decision-maker who aims to find out the truth from a suspect delegates to an interrogator with possibly different preferences. The ideal interrogator is always biased, sometimes nicer and sometimes tougher. The decision-maker can further improve by delegating to a nicer interrogator when the evidence is weak and to a tougher interrogator when the evidence is strong. Provided the decision-maker does not learn the details of the interrogation, she may then equivalently retain authority over decisions. Sequential delegation with leniency for perjury can implement the optimal mechanism with full commitment.

Keywords: bias, questioning, lie, confession, authority, evidence
JEL classifications: D82, D83, C72, K40

Alessandro Ispano: CY Cergy Paris Université, CNRS, THEMA, F-95000 Cergy, France, alessandro.ispano@gmail.com; Péter Vida: CY Cergy Paris Université, CNRS, THEMA, F-95000 Cergy, France, and Corvinus University of Budapest, Corvinus Institute for Advanced Studies, vidapet@gmail.com. We thank Attila Ambrus, Helmuts Āzacis, Françoise Forges, Takakazu Honryo, Joel Sobel, Egor Starkov and seminar participants at University of Graz theory seminar and CIAS workshop at Corvinus University of Budapest for useful comments. Financial support from Labex MME-DII and ANR grant 19-CE26-0010-03 is gratefully acknowledged.

1 Introduction

Should a mother who suspects her child to have disobeyed directly confront him to find out the true or let the gentler, softer, grandmother handle the situation? Likewise, once a suspect is brought in for questioning, who should proceed between an officer known for his toughness and propensity to incriminate at the first suspicion and one whose main concern is avoiding detaining an innocent? Or perhaps should the former only step in if the latter's attempt to elicit the true from the suspect is unsuccessful? Should these decisions depend on the strength of the incriminating evidence? And should the interrogation be secret or public? Should the preferences of law enforcers be fully aligned with the ones of society or differ?

This paper sheds light on all these issues building on the model of interrogations developed in Ispano and Vida (2021). There, we consider other design aspects, but we always maintain full alignment between the designer's and the interrogator's preferences. Here, we focus on the interrogator's preferences as design instrument, which we refer to as delegation. Preferences are parametrized by the relative cost of type II errors (e.g. exonerating a guilty) over type I errors (e.g. accusing an innocent). Thus, while both the designer and the interrogator aim to find out the true from the suspect, they may disagree on the decision to take when the suspect's status as guilty or innocent remains uncertain. Initially, we restrict our attention to one-shot communication from the suspect to the interrogator, who then takes a decision. We then consider more complex protocols of communication and allocation of authority over decisions.

Our first main insight is that the interrogator's ideal preferences from the point of view of the designer are always biased. In spite of the sub-optimality of the interrogator's decisions, some bias is desirable because of its impact on the informativeness of the interrogation. The direction of the interrogator's ideal bias may be towards accusation, i.e. a higher relative weight than the designer attached to type II errors, but also towards exoneration. In the former case, more guilties, i.e. even those who expect the incriminating evidence to be rather weak, confess instead of denying, since there is a cost from getting caught in a lie and the interrogator's decision upon an undetected lie becomes on average less favorable. In the latter, guilties resort to fewer lies, since smaller undetected lies suffice to be let go, so that innocents, who are always honest in equilibrium, are more easily set apart. This effect is akin to that of a stronger level of protection of the suspect's right to silence (Seidmann, 2005; Ispano and Vida, 2021) and, as a result, both the designer and the suspect benefit.

We then demonstrate how the accuracy of decisions further improves if the designer can condition the interrogator's bias on the strength of the evidence and, in particular, appoint

an interrogator more inclined towards accusation when the evidence is strong and towards exoneration when the evidence is weak. The suspect is kept in the dark in that he knows how the designer delegates but not the preferences of the actual interrogator, which would otherwise give away information about the evidence. The underlying mechanism is subtle and relies on a change in the shape of the lying strategy of guilties, who are induced to use smaller lies more often. Small lies are detected, hence punished, less often, which also means the decision upon an undetected lie is on average less favorable, reducing type II errors. The improvement from this conditional delegation policy is maximal when the preferences of the two types of interrogators become as extreme as possible, hence maximally biased towards exoneration and towards accusation provided such types are available. Also, when tailored appropriately, this policy is optimal among all one-shot delegation policies.

In addition to its effectiveness and intuitive appeal, this conditional delegation policy is immune to the designer's lack of commitment not to overrule the interrogator's decisions. Indeed, provided the designer only has the minimal knowledge required to implement the policy, i.e. she only knows to which interrogator she should delegate but not the exact evidence nor the suspect's message, she may equivalently retain authority over decisions and rely on the interrogator's recommendations. Besides, the game can be modified so that both the interrogator biased towards accusation and the one biased towards exoneration are present at the interrogation and provide recommendations, thereby providing a micro foundation for the well-known "good cop-bad cop" tactic in a fully rational framework without any reference to behavioral or psychological aspects. However, each of these constructions has some limitations. When a single interrogator is present, she should misrepresent her preferences not to give away information about the evidence. When both interrogators are present, the equilibrium is not immune to "neologisms", i.e. credible speeches one of the two interrogators would want to give to the designer to overturn the recommendation of the other. Regardless, the evidence may leak to the designer. Finally, if the interrogator maximally biased towards exoneration attaches even the tiniest disutility to letting a guilty go, the designer cannot attain her optimal payoff under full commitment.

To fix these issues, we provide a sequential delegation game that robustly implements the optimum involving the designer and an interrogator maximally biased towards accusation, again in the spirit of the good cop-bad cop tactic. They start the interrogation together and the suspect gives away some information anticipating if the evidence is weak relative to his claim the designer will let him go. When this is not the case, the designer "leaves the room" and the tough interrogator continues, which indirectly conveys information about the strength of the evidence the suspect. In the second round, a guilty suspect abandons his first-round lie

and confesses, thereby avoiding punishment, and an innocent keeps denying, which may lead to either accusation or exoneration depending on the strength of the evidence. The importance of the game lies in its information structure, but the interpretation need not be literal, e.g. the first round of the interrogation may be carried out by an unbiased police offer, the second round privately by a fully biased prosecutor, and the final decision by an unbiased judge.

Our results have clear normative implications for institutions who have eliciting information, in particular about one party's factual guilt, as one of their primary goals, most notably the judicial system, but also internal investigation committees in companies, anti-fraud authorities in a university, etc. Implications pertain to the definition of statuary objectives and procedures as well as to the appointment of agents based on their intrinsic preferences and the design of their incentive schemes. Even if extremely stylized, our model can rationalize recurrent traits of these institutions such as separation of roles, allocation of competence based on the severity of the case, information barriers and hierarchical talks. We single out four general observations.

- The presence of agents with biased preferences and whose discretion yield to ex-post suboptimal outcomes is not necessarily the symptom of a poorly managed institution. It is precisely the discretion that such agents enjoy in eliciting information and in taking or recommending decisions that can reduce errors.

- Diversity can be an asset. Thus, heterogeneity in agents' preferences, in particular with extreme biases, need not be the result of uncontrolled, possibly undesirable, self-selection into jobs. Promoting such heterogeneity can represent a purposeful managerial strategy.

- Transparency can hinder effective delegation. For instance, while desirable on other grounds, the use of one-way mirrors and the mandatory recording of interrogations may also undermine the credibility of society not to overturn law enforcers' decisions. The same is true for mandatory disclosure and sharing of the evidence within an organization.

- Keeping the party from which information must be elicited uncertain about the exact objectives of the agents he is facing and about who has real authority can be a key strategic asset.

Related literature. This paper lies at the intersection of law and economics and the literature on strategic communication and delegation. The law and economics literature on the judicial process generally takes as given the preferences of law enforcers, be them perfectly aligned with the ones of society (e.g. Grossman and Katz (1983), Reinganum (1988) and Seidmann (2005))

or possibly misaligned (e.g. Mialon (2005)). In the context of interrogations, this paper focuses on what the ideal misalignment is. The benefits of misalignment are due to its impact on the communication strategy of the player from which information must be elicited. This aspect differentiates this paper from the vast literature in which the principal delegates authority to an agent to benefit from his private information or his incentives to acquire it.[1]

Most closely related are the results on intermediation in communication of Dessein (2002) and Ivanov (2010). Both papers build on the seminal model of cheap-talk between a sender and a receiver (Crawford and Sobel, 1982) by interposing a strategic intermediary and investigate her ideal bias from the receiver's viewpoint.[2] In Dessein (2002) the receiver delegates to the intermediary both communication with the sender and decision-making, whereas in Ivanov (2010) she delegates communication but retains authority over decisions after hearing the intermediary's recommendation. Thus, as in this paper, the delegate's private information relative to the receiver obtains fully endogenously from communicating with the sender. However, the information and incentive structures are very different. In particular, in Dessein (2002) and Ivanov (2010), although biases can be small, large, or in opposite directions, there is no notion of tougher or nicer player. Moreover, the receiver has no private information relative to the sender. In our setting, by conditioning delegation on her evidence, the receiver can keep the sender uncertain about the intermediary's bias. Differently from Ivanov (2010) and Ambrus et al. (2013), in which the sender's and the designer's uncertainty stems only from randomization by the intermediary and suffices to reach the optimum, in our case the extra layer of uncertainty about the intermediary's preferences is key to discipline the suspect's behavior.

The decision-making procedure we consider is also different from those in the literature on delegation to teams. For example, Li and Suen (2004) consider simple unilateral acceptance or rejection procedures with two experts and ask whether the decision maker has an incentive to overrule decisions. In our case, the decision-maker chooses which interrogator's recommendation to accept based on the evidence. While different in the incentive and information structure and questions of interest, our paper is hence also related to the literature which studies how the allocation of authority among misaligned parties shapes communication and decisions in organizations.[3] Finally, our paper is related to the literature on the preferences of bureaucrats,

[1] See Holmstrom (1984) and Alonso and Matouschek (2008) as examples of the former and Szalay (2005), Deimen and Szalay (2019), and Ball and Gao (2020) as examples of the latter. Central questions in this literature are whether the principal should set limit to the agent's discretion and whether delegation is superior to communication from the agent. The idea of delegation as a commitment device in games dates back to Schelling (1956). For a review using a managerial perspective, see Sengul et al. (2012).

[2] See also Ambrus et al. (2013) and Chen and Gordon (2015), who consider respectively several intermediaries and more general preferences.

[3] See for instance Alonso et al. (2008) and Chakraborty and Yılmaz (2017).

police officers in particular, which however typically focuses on different traits and roles.[4] For example, in Prendergast (2007) bureaucrats' bias is desirable because it incentivizes them to exert monitoring effort. Also, while "bifurcation" of bureaucrats' preferences towards extreme biases is the result of self-selection, in our setting it is an optimal choice of the designer.

2 Model

We model the interrogation as a game of two-sided incomplete information between a suspect (S, he) and a law enforcer (R, she). At the initial stage, S's private information, or type, y and R's private information, or evidence, z are drawn uniformly from $\{(y,z) \in [0,1]^2 : y < z\}$. S is guilty if $y < t$ and innocent otherwise, with $t \in (0,1)$ exogenously given. The evidence z proves that $y < z$, the lower the z the stronger the evidence. The evidence is conclusive if $z \leq t$ and inconclusive if $z > t$. S's type y, in addition to determine whether he is innocent or guilty, determines the strength of the evidence he expects R to possess, which is stronger the lower the y. S then sends R a message $m \in \mathcal{M} = [0,1]$, interpreted as a literal claim about his type y, so that S confesses when $m < t$ and denies when $m \geq t$. Finally, R takes an action $a \in \{0,1\}$, e.g. accuses S ($a = 0$) or exonerates him ($a = 1$), and payoffs realize. R aims to minimize a weighted sum of type I (accuse an innocent) and type II errors (exonerate a guilty):

$$\alpha\, a \mathbb{1}_{y<t} + (1-\alpha)\,(1-a)\,\mathbb{1}_{y\geq t}, \tag{1}$$

where $\mathbb{1}$ is the indicator function and $\alpha \in (0,1)$ is the weight of type II errors. S's payoff is $a - b\mathbb{1}_{m\geq z}$, i.e. his utility from being accused and exonerated is normalized to 0 and 1 respectively, and he also suffers some cost $b > 0$ when he is caught in a lie ($m \geq z$). For further details, interpretations and extensions of the model, we refer to Ispano and Vida (2021). In particular, we explain how the incentive structure is ultimately equivalent to one in which S incurs no cost b but he enjoys some leniency $b/(1+b) \in (0,1)$ for confessing.

2.1 Equilibrium without delegation

Following Ispano and Vida (2021), we restrict our attention to pure strategy weak perfect Bayesian equilibria in which innocent types and confessors are honest, i.e. types $y \geq t$ send $m = y$ and types $y < t$ who send $m < t$ choose $m = y$. W.l.o.g., a pure strategy of S can be

[4] For instance, see Friebel et al. (2019) for an experiment on police officers' attitudes towards enforcing cooperation.

fully described by a partition of the set of guilty types into a possibly empty set of types who confess ($y < y_c$ with $y_c \in [0, t)$) and into those who lie ($y \in [y_c, t)$) and by a surjective, strictly increasing, lying function $\boldsymbol{\ell} : [y_c, t) \to [t, \bar{y})$ which associates to each type $y \in [y_c, t)$ a denying message $\ell(y) \in [t, \bar{y})$, where $t < \bar{y} < 1$. R's belief system, which assigns a probability to the innocence of S given m and z with $z > m \geq t$, is consistent with (a generalized version of) Bayes' rule. That is, it is equal to $1/(1 + \ell^{-1'}(m))$, which importantly is independent from z, whenever $\boldsymbol{\ell}$ is differentiable, i.e., for almost every m in the range of $\boldsymbol{\ell}$, and arbitrary otherwise unless the message is only sent by a guilty or by an innocent type. In the latter two cases the belief is 0 and 1 respectively and so are the corresponding sequentially rational actions of R. When $z \leq m$ the belief and the action of R must be 0 because S is caught in a lie, i.e. his message is inconsistent with the evidence, and innocents are honest. We are left with specifying R's actions in the range of $\boldsymbol{\ell}$ when $z > m$. W.l.o.g., we concentrate on cut-off strategies of R of the form $a(m, z) = 1$ if $z \geq \bar{z}(m)$ and 0 otherwise specified by some $\bar{\boldsymbol{z}} : [t, \bar{y}) \to [t, 1]$. Of course, in equilibrium, the decision of R must be sequentially rational given her belief. Likewise, the message sent by any type of S, including honest types, must be sequentially rational given R's reaction.

Proposition 0 (Unique Equilibrium)**.** *For any $\alpha \in (0, 1)$, there is a unique equilibrium in which:*

(i) low guilty types (if any) confess and high guilty types lie;

(ii) liars lie according to a strictly increasing lying function and pool with low innocent types, while sufficiently high innocent types separate;[5]

(iii) upon a pooling message and not catching S in a lie R is indifferent between exonerating or accusing S;

(iv) all guilty types who do not confess are indifferent with respect to any of the equilibrium lies.

Proof. See Ispano and Vida (2021). □

The equilibrium values are uniquely determined as follows. If $b \leq \frac{1-t-\alpha}{t}$ then $y_c = 0$, i.e. there are no confessors, and $\bar{y} = \frac{t}{1-\alpha}$. If instead $b > \frac{1-t-\alpha}{t}$ then $y_c = \frac{(1+b)t-(1-\alpha)}{b+\alpha} > 0$, i.e. there

[5]Points i and ii follow directly from the definition of the strategies. In Ispano and Vida (2021) we show that when allowing for larger strategy spaces, for example arbitrary partitions or requiring only measurability of $\boldsymbol{\ell}$, any other equilibria has eventually the same structure and is payoff equivalent ex-ante and ex-post for both S and R.

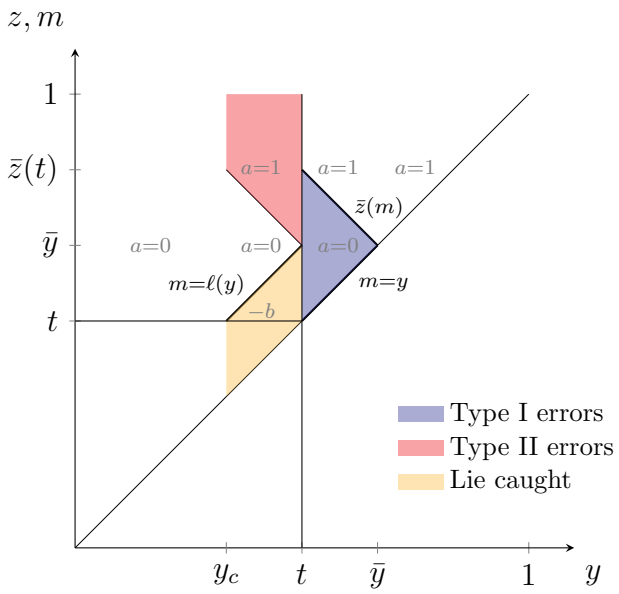


Figure 1 Equilibrium payoffs
($t = 1/2$, $b = 1$, $\alpha = 1/2$)

are confessors, and $\bar{y} = \frac{\alpha+bt}{\alpha+b}$. In both cases $\ell(y) = t + \frac{\alpha}{1-\alpha}(y - y_c)$ and $\bar{z}(m) = \bar{y} + b(\bar{y} - m)$. Hence, after any message $m \in [t, \bar{y})$, whenever $m < z$ R's belief about S's innocence is α. Furthermore, R's ex-ante expected loss is proportional to

$$(1-\alpha)\underbrace{\int_t^{\bar{y}} (\bar{z}(y) - y)\mathrm{d}y}_{\text{type I errors}} + \alpha \underbrace{\int_{y_c}^{t} (1 - \bar{z}(\ell(y)))\,\mathrm{d}y}_{\text{type II errors}} \tag{2}$$

The equilibrium and the corresponding type I and type II errors are demonstrated in figure 1, taken from Ispano and Vida (2021).

Note the proposition describes the equilibrium only for $\alpha \in (0, 1)$. Henceforth, unless specified otherwise, when the interrogation is delegated to an interrogator who has extreme preferences, i.e. who only cares about type I or type II errors, we consider the limit of the equilibria as α goes respectively to 0 and 1.[6] The limit of these equilibria is indeed an equilibrium of the limit game, even though there may be others that are not payoff equivalent.

3 Simple delegation

Suppose R can choose to delegate the interrogation to an interrogator who also observes evidence z and whose loss is still given by equation (1) but with an arbitrary, possibly different

[6] In the limit as α goes to 0 the distribution of lies is all concentrated at t and, upon observing $m = t$ and not catching S in a lie, the interrogator sometimes exonerates S even though she is sure he is guilty. Instead, in the limit as α goes to 1 the measure of liars is zero and upon observing a message in the lying region, the interrogator sometimes accuses S even though she is sure he is innocent.

relative weight for type II errors. The interrogator's preferences are known to S and the interrogation then plays out as in the baseline model, except that it is now the interrogator who takes decisions. Formally, R, with preference $\alpha \in (0,1)$, wishes to choose an $\alpha^\star \in [0,1]$ such that her ex-ante expected loss (equation (2), still evaluated according to her preference α) is minimized when the resulting equilibrium objects are determined by α^*.

Proposition 1 (Simple delegation)**.** *Let $\alpha^\star$ denote the preference of R's optimal interrogator.*

(i) The optimal interrogator is always biased ($\alpha^\star \neq \alpha$).

(ii) When some types confess absent delegation ($b > \frac{1-t-\alpha}{t}$), the optimal interrogator is always tougher than R ($\alpha^\star > \alpha$), and maximally tough ($\alpha^\star = 1$) if and only if R is sufficiently tough ($\alpha \geq 2/3$).

(iii) When no type confesses absent delegation ($b \leq \frac{1-t-\alpha}{t}$), the optimal interrogator is nicer than R ($\alpha^\star < \alpha$) provided R is sufficiently nice ($\alpha < \frac{2(1-t-bt)}{3+t}$), but never maximally nice ($\alpha^\star > 0$).

Proof. See section A.1 in the appendix. □

R's choice to delegate to a tougher interrogator affects R's expected loss in three ways. First, it yields to suboptimal decisions biased towards accusation. Second, it disciplines S to favor confession over lying. Third, it induces types who still elect not to confess to also use bigger lies, i.e. the lying region increases. Starting from a situation in which the set of confessors would be non-empty without delegation, the informational benefit of increased confession at least initially dominates the two other negative effects. Thus, in this case R always finds it optimal to delegate to a tougher interrogator. In particular, the interrogator should be maximally biased towards minimizing type II errors if exonerating a guilty is already rather costly for R. Instead, when given R's preferences the set of confessors would be empty without delegation, the minimal interrogator's toughness required to benefit from increased confession may be too far off. In this case, R prefers a nicer interrogator because, in spite of the suboptimal decisions biased towards exoneration, the lying region decreases, enhancing separation of innocents and guilties. Since the interrogator's decisions become on average more favorable, i.e. $\bar{z}(m)$ decreases, also S is better off.

4 One-shot conditional delegation

4.1 Intuitive delegation

Suppose now R can condition the interrogator's preferences on the strength of the evidence. Formally, R commits to a delegation policy $\boldsymbol{\alpha} : [0,1] \to [0,1]$ which determines the preference parameter $\alpha(z)$ of the actual interrogator given evidence z. S observes the policy but not the preference parameter of the actual interrogator, which may otherwise convey information about z (see more on this in section 5.3). Then the interrogation unfolds as in the baseline model.

To demonstrate how conditional delegation outperforms simple delegation, we concentrate on cutoff policies, which we refer to as *intuitive*, that prescribe delegating to a nicer interrogator when the evidence is sufficiently weak and to a tougher one otherwise. We demonstrate by an elaborated example here below how the maximal improvement obtains by delegating to the nicest and toughest interrogator and why this improvement is the most effective in general, i.e. it is optimal among all possible one-shot conditional delegation policies one can think of, including those, possibly random policies, which condition the interrogator's preferences also on the message of S. Nothing is special about the example, so that the arguments apply to any parameter combination. We summarize these insights in the following proposition.

Proposition 2 (Improvements by intuitive delegation)**.**

(i) (Nice-weak & tough-strong) *For any simple delegation policy that is not extreme, i.e.* $\boldsymbol{\alpha} = \alpha_{const} \in (0,1)$*, there exists a strictly loss-reducing cut-off policy that prescribes delegating to a nicer interrogator, i.e. with preference* $\alpha_n < \alpha_{const}$*, when the evidence is sufficiently weak, i.e. when* $z \geq \tilde{z}$ *for some* $\tilde{z}$*, and to a tougher interrogator, i.e. with preference* $\alpha_\tau > \alpha_{const}$*, otherwise.*

(ii) (Robustness to R's preference) *This policy can always be chosen so as to reduce type II errors while leaving type I errors unaffected (relative to those under* α_{const}*) and hence it is preferred by* R *independently of her actual preferences.*

(iii) (Extremity and optimality) *The loss reduction from this policy is maximal when the preferences of the nicer and the tougher interrogator are as extreme as possible, i.e.* $\alpha_n = 0$ *and* $\alpha_\tau = 1$*. Moreover, the optimal cut-off policy, described by a* $\tilde{z}^*(\alpha)$ *such that* $\alpha_n = 0$ *when* $z \geq \tilde{z}^*(\alpha)$ *and* $\alpha_\tau = 1$ *otherwise, is optimal among all possible one-shot conditional delegation policies.*

(iv) (Always a grain of niceness and toughness) *Simple delegation to an interrogator with extreme preferences is never an optimal policy, i.e.* $\tilde{z}^*(\alpha) \in (t, 1)$ *for any* $\alpha \in (0, 1)$.

4.1.1 The example

Fix $t = 1/2$, $b = 1$ and $\boldsymbol{\alpha} = \alpha = 1/2$, constant. Using proposition 0, in equilibrium without delegation $y_c = 1/3$ and $\bar{y} = 2/3$, so that the measure of liars and of lies sent are both equal to $1/6$. Upon a pooling denying message not disproven by the evidence, the interrogator is indifferent and chooses $a = 1$ when $m \geq \bar{z}(m) = 4/3 - m$ and $a = 0$ otherwise. The equilibrium and the resulting type I and type II errors are displayed in figure 1.

Suppose instead R delegates to an interrogator with preference parameter $\alpha_n = 1/4$ when $z \geq \tilde{z} = 19/24$ and to an interrogator with preference parameter $\alpha_\tau = 3/4$ when $z < \tilde{z}$. Suppose also (we will ensure that this is indeed sequentially rational) that in her respective region of competence each interrogator still follows the same cutoff strategy $\bar{z}(m)$ as in the original equilibrium. Then, S's incentives are completely unaffected, so that y_c is the same and, in particular, liars are still indifferent to any lie in $[t, \bar{y})$. It is then possible to construct a lying function with image $[t, \bar{y})$ such that each interrogator finds it optimal to follow $\bar{z}(m)$.

To see this, let us denote by $\tilde{m} = 13/24$ the message such that $\tilde{z} = \bar{z}(\tilde{m})$, as represented in figure 2. Looking from the perspective of the horizontal axis after projecting $\tilde{m}$ onto it, when $m \in [t, \tilde{m}]$ the nice interrogator takes both action $a = 0$ and $a = 1$ based on z. This is only possible if she is indifferent which will be ensured by the appropriate shape of the lying function. In turn, this implies that the tough interrogator indeed finds it strictly optimal to always choose $a = 0$. Likewise, when $m \in (\tilde{m}, \bar{y})$, the tough interrogator takes both action $a = 0$ and $a = 1$, which again is only possible if she is indifferent as will be ensured by the appropriate shape of the lying function. The nice interrogator then indeed finds it strictly optimal to always choose $a = 1$. Easy calculations show that these two indifference conditions hold for the following lying function

$$\ell(y) = \begin{cases} 7/18 + 1/3y & \text{if } y \in [y_c, \tilde{y}) \\ 17/18 - 3y & \text{if } y \in [\tilde{y}, t)\,, \end{cases}$$

where $\tilde{y} = 11/24$ is the guilty type who sends message $\tilde{m}$ (at $\tilde{m}$, the belief of α_n should be no less than $1/4$ and the belief of α_τ no more than $3/4$, e.g. they can share any belief in this region).

This lying function is depicted in red in figure 2. A comparison with figure 1 illustrates how type I errors, as well as type II errors for $z \geq \bar{z}(t) = 5/6$, remain unaffected relative original equilibrium. Instead, type II errors for $z \in (t, \bar{z}(t))$ decrease of the green region, whose area

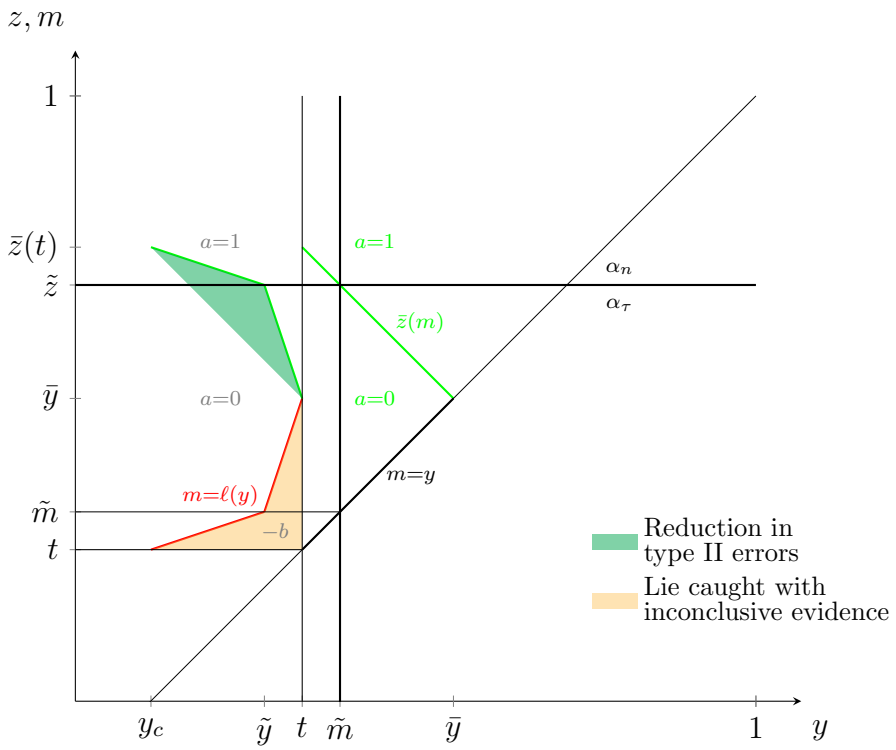


Figure 2 The benefits of conditional delegation

has size 1/144. The change in the lying function induced by conditional delegation shifts the distribution of lies towards lower messages. As a result, S is caught in a lie and hence punished less often, so that R also chooses $a = 1$ less often.

From these arguments, one can first of all see that the chosen delegation policy is actually the optimal one among the ones $\boldsymbol{\alpha} : [0, 1] \to [1/4, 3/4]$ that leave y_c unaffected. The same is true if R can randomize or also condition delegation on the received message. Indeed, R aims to make the lying function as flat as possible before the kink and as steep as possible after the kink - this is in fact how we computed $\tilde{z}$ in the first place. Besides, as the preferences of the nicer and tougher interrogator gets more extreme, lies get more and more concentrated around t. In the limit, these type II errors entirely disappear since lies are never caught by inconclusive evidence (yet all lies will still be caught with conclusive evidence, so that the $-b$ area in figure 1 does not disappear completely, as it will be case under optimal delegation, discussed in section 5.3). In this limit case, upon observing $\tilde{m} = t$, the nice interrogator is now (almost) sure to face a guilty type (because all the guilties lie to t) but, as type II errors yield her no disutility, her indifference condition is preserved.

The only special feature of the example is that the starting preference of the interrogator is not extreme. Proposition 1 already rules out that simple delegation to a maximally nice interrogator is optimal. However, nor is simple delegation to a maximally tough interrogator. Indeed, tedious but straightforward calculation shows that the optimal cut-off delegation policy

$z^*(\alpha, t, b)$ is $\frac{-\alpha bt+2\alpha b+2\alpha+bt}{\alpha b+2\alpha+b}$ (and there are confessors) when $t \geq \bar{t} \in [\frac{1-\alpha}{(1+\alpha)(1+b)}, \frac{1-\alpha}{1+b}]$ and it is $\frac{t(1+\alpha b)}{1-\alpha}$ (and there are no confessors) when $t \leq \bar{t}$. Since it is always the case that $z^*(\alpha, t, b) \in (t, 1)$, point iv of proposition 2 obtains.

4.2 Alternative delegation policies

We note that, if the choice of the actual interrogator could be made ex-post, the same benefit of intuitive delegation could be reached by conditioning delegation exclusively on the messages of S. Indeed, from the example, it is apparent that exactly the same equilibrium would arise by delegating to α_n when $m \leq \tilde{m}$ and to α_τ when $m > \tilde{m}$. Likewise, as we explain in section A.3 of the appendix, the same effect can obtain by *reverse delegation*, i.e. appointing the tough interrogator when the evidence is weak and the nice one when the evidence is strong, and by *random delegation*, i.e. choosing the interrogator randomly and unconditionally. However, in all of these constructions, in contrast to intuitive conditional delegation, R would want to overrule the decision of the interrogator (see section 5.1).

5 Discussion

In this section we consider several extensions. First, we investigate the commitment problem of R. Namely, we ask whether and when R has an incentive to overrule the decision of the actual interrogator under intuitive delegation and alternative policies. Second, we consider a good cop-bad cop scenario in which both the nice and the tough interrogator are present and compare it to intuitive delegation. Finally, we describe a game using sequential conditional delegation and discretionary punishment of lies that can robustly implement the second-best achievable payoff for R in such a way that her commitment problem is the least severe.

5.1 Retaining authority

We now show that R can implement the optimal intuitive delegation policy described in section 4.1 in such a way that she, herself, makes the final decision after collecting garbled information from the actual interrogator. That is, she delegates only the communication part of the interrogation, she then asks a recommendation from the actual interrogator, i.e. from the nice interrogator α_n if the evidence is weak and from the tough interrogator α_τ when the evidence is strong, and finally she takes a decision. Under the assumption that R knows, or learns, only whether z is weaker or stronger than the optimal cut-off $\tilde{z}^*(\alpha)$ (or equivalently, that

the recommendation comes from the nice or the tough interrogator, respectively) but not the actual value of z nor S's message, she finds it sequentially rational to follow the recommendation. In turn, it is sequentially rational for the actual interrogator to recommend the decision that she would take under full delegation and hence S's equilibrium strategy remains unchanged.[7]

Proposition 3 (Authority and intuitive delegation)**.** *Consider the optimal intuitive delegation policy at proposition 2. If R knows only whether the evidence is stronger or weaker than $\tilde{z}^*(\alpha)$, then she has no incentives to overrule the decision of the actual interrogator. In particular, R learns from α_n only whether S confessed, but not whether an S who did not confess is innocent or guilty. Moreover, R does not learn from α_τ whether S confessed but sometimes she learns that S is innocent.*

Proof. See section A.2 in the appendix. □

We have mentioned in section 4.2 that R can reach the optimal one-shot delegation not only by means of the intuitive policy but also by conditioning on the message of S, by reverse delegation, and by random delegation. A closer look at the structure of the equilibria in these games reveals that the corresponding versions of proposition 3 fail in all these cases. This failure is obvious when delegation is based on S's message. In this case α_n recommends a decision only if the message of S is in $[0, t]$ and α_τ recommends otherwise.[8] Then, in equilibrium, R always wants to exonerate S if α_τ accuses him and always wants to accuse S when α_n exonerates him. In case of reverse delegation, communication with α_τ poses the same incentive constraint as in the case of proposition 3. However, the incentive constraint when communicating with α_n becomes tighter, i.e. it will be harder, and sometimes plainly impossible, to accept α_n's recommendation to exonerate S. In case of random delegation, both constraints become tighter.

These observations demonstrate that when R cannot commit not to overrule decisions the problem is the least severe under intuitive delegation. Besides, R should learn only very coarse information about the evidence. And finally, R should be able to commit not to learn S's actual message. This last point suggests that, to avoid potential information leakages, only the appointed interrogator should be present at the interrogation even when both types of interrogators can assist (see section 5.2 for further details). This solution, however, poses another problem since, as mentioned before, the actual interrogator should not provide information about

[7] By construction, on the equilibrium path α_n never detects a lie and α_τ detects lies only with conclusive evidence. Hence, R can learn from α_τ that a lie was detected, R's incentives to follow α_τ's recommendation to accuse remains unchanged and R could also take care of the punishment $-b$.

[8] If α_n and α_τ were in $(0, 1)$ then the structure of the equilibrium would be the same. α_n recommends a decision after weak denying messages (i.e. when $m \leq \tilde{m}$ for some $\tilde{m} > t$) and otherwise α_τ does so (see section 4.1.1 in the appendix).

the evidence to the suspect. Therefore, she should be able to conceal her true preferences. Or, equivalently, at least one of the two types of interrogators should pretend and behave as the other.

5.2 Delegating to teams, good cop-bad cop

In presenting intuitive delegation, we have assumed that R delegates the interrogation to either the nice interrogator α_n or the tough interrogator α_τ depending on the strength of the evidence and receives a recommendation only from either of the two. Alternatively, in the spirit of the good cop-bad cop tactic, one can imagine a scenario in which both α_n and α_τ are present at the interrogation. It then becomes irrelevant whether S can distinguish them and therefore they do not have to pretend as in the case of intuitive delegation. On the contrary, their natural behavior can make it visible for the suspect that he really faces two different types and that R has indeed delegated as anticipated. Assuming R retains authority over decisions and receives a recommendation from each of them, we can ask whether the equilibrium construction carries through and, in particular, how R would act in case of disagreement.

Consider the communication strategies of α_n and α_τ as in the equilibrium of proposition 3 and complete their strategies when they are not appointed as follows. We specify that α_n always recommends exoneration when the evidence is strong even if she observed a confession and α_τ always recommends accusation when the evidence is weak. In this way α_n and α_τ "mute" themselves in equilibrium when the evidence is strong or weak respectively, and do not provide additional information for R beyond what R learns from α_n and α_τ when the evidence is weak or strong respectively. Hence, R's incentives do not change relative to the scenario described in proposition 3. To complete the description of the equilibrium, R follows the recommendations if they agree, otherwise R follows the recommendation of α_n or α_τ in their respective region of competence.

Note, however, that such equilibrium construction assumes the two interrogators can recommend only exoneration and accusation (and of course warning R about the punishment when needed). Namely, their message space is restricted to $\{0, 1\}$. Yet, when the evidence is strong and S's message is $m > t$, α_n would like to reveal this information, i.e. make a speech along the lines of "Look, S's message was actually strictly larger than t". A similar argument applies to α_τ when the evidence is weak and S's message is t. Using the terminology of Farrell (1993), the equilibrium is not neologism-proof.

5.3 Optimal delegation

In this section, first we recall from Ispano and Vida (2021) what the best payoff of R can achieve is by fully committing to a mechanism is, which also assumes that R cannot misrepresent her evidence z. This setup is called arbitration in Goltsman et al. (2009). Then, we show that this payoff can be achieved by intuitive delegation at the limit as the preference of the nice and tough interrogator are extreme, i.e respectively $\alpha_n = 0$ and $\alpha_\tau = 1$. However, differently from the equilibria at propositions 2 and 3, this equilibrium is not robust to small perturbations of $\alpha_n = 0$ (all of these equilibria are robust to perturbations of $\alpha_\tau = 1$). To fix this problem and those mentioned above in sections 5.1 and 5.2, we consider a game in which R first interrogates S together with α_τ and then, given her evidence and S's message, she can decide to exonerate S or to leave the room and delegate the continuation of the interrogation to α_τ, even in case R has detected a lie. α_τ then interrogates S again and takes a decision (or equivalently makes a recommendation to R) given the evidence and the two messages of S. This equilibrium is robust to small perturbations of $\alpha_\tau = 1$ in that R's payoff gets arbitrarily close to her arbitration one if α_τ is sufficiently close to 1.

To this end, w.l.o.g., consider cut-off mechanisms $\hat{\boldsymbol{z}} : [0,1] \to [0,1]$ which specify for each message y a cutoff level $\hat{z}(y) \in [y,1]$ such that $a(y,z) = 1$ if and only if $z \geq \hat{z}(y)$. The optimal direct mechanism $\hat{\boldsymbol{z}}^\star$ minimizes

$$\alpha \int_0^t (1 - \hat{z}(y))\mathrm{d}y + (1-\alpha) \int_t^1 (\hat{z}(y) - y)\,\mathrm{d}y \tag{3}$$

subject to the constraint that each type finds it weakly optimal to be honest and not lie upward, i.e. for every $y, y' \in [0,1]$ such that $y < y'$

$$1 - \hat{z}(y) \geq 1 - \hat{z}(y') - b\,(y' - y)\,. \tag{4}$$

It turns out (see proposition 6 in Ispano and Vida (2021)) that $\hat{\boldsymbol{z}}^\star$ is the extension of $\bar{\boldsymbol{z}}$ and it is equal to $\bar{\boldsymbol{z}}$ for $y \in [y_c, \bar{y}]$, while its value is 1 for $y < y_c$, i.e. these types are always accused, and y for $y \in [\bar{y}, 1]$, i.e. these types are always exonerated. Thus, under the optimal mechanism S's payoff and type I errors are just as in the equilibrium at proposition 0. Hence, the optimal mechanism can be easily implemented with intuitive delegation with cut-off $\bar{z}(t)$ as an equilibrium of the limit game where $\alpha_n = 0$ and $\alpha_\tau = 1$. In this equilibrium, each type of S is honest and the actual interrogator follows cut-off strategy $\hat{\boldsymbol{z}}^\star$. Moreover, R has no incentives to overrule the decision of the actual interrogator even if she knows the exact value of z. Hence,

the equilibrium payoff can be implemented in a game where R makes the decision following the recommendation of the actual interrogator. This equilibrium, however, is not robust to small perturbations of α_n, i.e. it is not a limit of equilibria as α_n goes to zero.

To dispense with this problem and with those mentioned in sections 5.1 and 5.2, consider now the following natural game with three players, S, R and α_τ, where $\alpha_\tau = 1$ is a maximally tough interrogator:

- **Stage 0** S and R observe their private information as in the baseline model. Additionally, α_τ also observes R's private information;
- **Stage 1** R and α_τ interrogate S, i.e., S sends them a public message $m \in \mathcal{M}$;
- **Stage 2** based on S's message m and the evidence z, R can either take a decision $a \in \{0, 1\}$, in which case the game ends and payoffs realize as in the baseline model, or choose to delegate the continuation of the interrogation to α_τ, so that stage 3 is reached;
- **Stage 3** α_τ interrogates S again by specifying a set of messages $\mathcal{M}_\tau \subseteq \mathcal{M}$ from which S can send a new message m_2;[9]
- **Stage 4** based on S's new message m_2 as well as on z and m, α_τ makes a recommendation $a_2 \in \{0, 1\}$ to R;
- **Stage 5** based on S's message m, on the evidence z and a_2, R takes an action, the game ends and payoffs realize as in the baseline model.

Proposition 4 (Implementation without commitment)**.** *There is an equilibrium of this game in which R's and S's expected payoffs are as in the optimal mechanism $\hat{\boldsymbol{z}}^\star$ and S's behavioral strategy in stage 1 is as in the equilibrium at proposition 0. The equilibrium is robust to small perturbations of α_τ.*

Proof. See section A.4 in the appendix. □

In stage 1, S uses the equilibrium strategy described in proposition 0. When S's message in separating, R immediately takes the correct action. When S's message in not separating, R exonerates S if the evidence is weak relative to the received message, i.e. if $z \geq Z(m)$, and otherwise she delegates the continuation of the interrogation to α_τ and commits not to eavesdrop the conversation between α_τ and S. In particular, $m < Z(m)$, i.e. a liar is never exonerated

[9] This modeling choice is just for simplicity. S could equivalently send any arbitrary message. Then S would get $-b$ if caught in a lie as in the baseline model. Additionally, α_τ could just impose a punishment $-b$ in case she does not "get an answer to her question".

when caught ($z \leq m$), and $Z(m)$ is decreasing, i.e. higher messages require stronger evidence for the interrogation to continue. It is the increasing chance of being exonerated that allows screening among guilty types unwilling to confess in stage 1. R finds it optimal not to deviate from her delegation policy due to the disciplining off the equilibrium path behavior of α_τ, who would then recommend accusation of S with probability one. When the interrogation continues, α_τ chooses $\mathcal{M}_\tau = \{\ell^{-1}(m), m\}$, i.e. asks S the question: "Are you m or $\ell^{-1}(m)$?". Guilty type $\ell^{-1}(m)$, who sent the pooling message m, hence gets a second chance to confess. The appropriate choice of $Z(m)$ now induces type $\ell^{-1}(m)$ to do so, as he learned that the evidence is strong from the fact that the interrogation continued without R. In order not to leave to S any unnecessary surplus and to reach payoffs as in the optimal mechanism, $Z(m)$ must be chosen to make type $\ell^{-1}(m)$ exactly indifferent between confessing and sticking to his stage 1 story m. Equilibrium lies in stage 1 are hence forgiven. Instead, innocent type m sticks to his stage 1 story, i.e. he sends $m_2 = m$, and α_τ may either recommend to accuse him or to exonerate him depending on the evidence. In equilibrium R follows the recommendation of α_τ.[10]

Thus, this sequential delegation game can solve most of the problems mentioned above: R can perfectly observe the evidence, no interrogator has to misrepresent her preferences, and the construction easily adjusts if α_τ attaches some small disutility to accusing an innocent. The unique commitment it imposes on R is not to eavesdrop the conversation between S and α_τ. While the structure of the game is natural, it is also rather articulated. Therefore, R may find it harder to commit to it and make it visible to S in certain settings, e.g. situations which are unfamiliar to S. This is why considering more straightforward institutions, such as simple or intuitive delegation, is also important. Besides, a comparison with the implementation of the optimal mechanism described in Ispano and Vida (2021) illustrates how sequential delegation can substitute for direct disclosure of information about the evidence to the suspect when this possibility is not available to the designer, e.g. because the evidence is soft information or classified.

References

Alonso, Ricardo and Niko Matouschek, "Optimal delegation," *The Review of Economic*

[10]Notice that on the equilibrium path α_τ will know that an S who is sticking to his story is surely innocent. Nonetheless, as α_τ is maximally tough she is indifferent between accusing S or exonerating him. If there is no access to a maximally tough interrogator, the equilibrium above obtains in the limit as α_τ gets tougher and tougher. The difference is that, to make α_τ indifferent, guilty types must now confess randomly with a probability approaching 1 as α_τ's toughness becomes maximal. We also note that there is an inverse game in which a maximally nice interrogator $\alpha_n = 0$ replaces α_τ and an associated inverse equilibrium implementing the same payoff. This equilibrium, however, is again not robust to small perturbations of α_n.

Studies, 2008, *75* (1), 259–293.

_ , **Wouter Dessein, and Niko Matouschek**, "When does coordination require centralization?," *American Economic Review*, 2008, *98* (1), 145–79.

Ambrus, Attila, Eduardo M. Azevedo, and Yuichiro Kamada, "Hierarchical cheap talk," *Theoretical Economics*, 2013, *8* (1), 233–261.

Ball, Ian and Xin Gao, "Benefiting from Bias," *Working paper*, 2020.

Chakraborty, Archishman and Bilge Yılmaz, "Authority, consensus, and governance," *The Review of Financial Studies*, 2017, *30* (12), 4267–4316.

Chen, Ying and Sidartha Gordon, "Information transmission in nested sender–receiver games," *Economic Theory*, 2015, *58* (3), 543–569.

Crawford, Vincent P. and Joel Sobel, "Strategic Information Transmission," *Econometrica*, 1982, *50* (6), pp. 1431–1451.

Deimen, Inga and Dezső Szalay, "Delegated expertise, authority, and communication," *American Economic Review*, 2019, *109* (4), 1349–74.

Dessein, Wouter, "Authority and communication in organizations," *The Review of Economic Studies*, 2002, *69* (4), 811–838.

Farrell, Joseph, "Meaning and credibility in cheap-talk games," *Games and Economic Behavior*, 1993, *5* (4), 514–531.

Friebel, Guido, Michael Kosfeld, and Gerd Thielmann, "Trust the police? Self-selection of motivated agents into the German police force," *American Economic Journal: Microeconomics*, 2019, *11* (4), 59–78.

Goltsman, Maria, Johannes Hörner, Gregory Pavlov, and Francesco Squintani, "Mediation, arbitration and negotiation," *Journal of Economic Theory*, 2009, *144* (4), 1397–1420.

Grossman, Gene M and Michael L Katz, "Plea bargaining and social welfare," *The American Economic Review*, 1983, *73* (4), 749–757.

Holmstrom, Bengt, "On the theory of delegation," in M. Boyer and K. Richard, eds., *Bayesian Models in Economic Theory*, North-Holland, 1984, pp. 115–141.

Ispano, Alessandro and Peter Vida, "Designing Interrogations," *Working paper*, 2021.

Ivanov, Maxim, "Communication via a strategic mediator," *Journal of Economic Theory*, 2010, *145* (2), 869–884.

Li, Hao and Wing Suen, "Delegating decisions to experts," *Journal of Political Economy*, 2004, *112* (S1), S311–S335.

Mialon, Hugo M, "An economic theory of the fifth amendment," *Rand Journal of Economics*, 2005, pp. 833–848.

Prendergast, Canice, "The Motivation and Bias of Bureaucrats," *The American Economic Review*, 2007, *97* (1), pp. 180–196.

Reinganum, Jennifer F, "Plea bargaining and prosecutorial discretion," *The American Economic Review*, 1988, pp. 713–728.

Schelling, Thomas C, "An essay on bargaining," *The American Economic Review*, 1956, *46* (3), 281–306.

Seidmann, Daniel J, "The effects of a right to silence," *The Review of Economic Studies*, 2005, *72* (2), 593–614.

Sengul, Metin, Javier Gimeno, and Jay Dial, "Strategic delegation: A review, theoretical integration, and research agenda," *Journal of Management*, 2012, *38* (1), 375–414.

Szalay, Dezsö, "The economics of clear advice and extreme options," *The Review of Economic Studies*, 2005, *72* (4), 1173–1198.

A Appendix

A.1 Proof of proposition 1

Throughout, let α_0 and α denote the preference parameter of R and of the interrogator, respectively. For any $\alpha \in (0,1)$, the equilibrium is described in section 2.1 (our analysis allows for $\alpha = 0$ and $\alpha = 1$ as limit cases given that R's expected loss varies continuously with the choice of α) and R's expected loss is now

$$E(\alpha) = (1-\alpha_0)\int_t^{\bar{y}} (\bar{z}(y) - y)\mathrm{d}y + \alpha_0 \int_{y_c}^{t} (1 - \bar{z}(\ell(y)))\,\mathrm{d}y. \tag{5}$$

The equilibrium expressions for y_c, $\bar{y}$, $\bar{z}(y)$ and $\ell(y)$, which all depend on α, differ depending on whether there are confessors in equilibrium, i.e. on whether $\alpha > \bar{\alpha} \equiv \max\{1-(1+b)t, 0\}$. Letting the subscripts c and nc indicate respectively the region with and without confessors throughout, we solve for R's optimal choices in the two regions, denoted respectively $\alpha_c^\star$ and $\alpha_{nc}^\star$, and then compare $E(\alpha_c^\star)$ and $E(\alpha_{nc}^\star)$ (keeping in mind that if $\bar{\alpha} = 0$, only the first case is possible).

There are confessors. When $\alpha > \bar{\alpha}$, replacing equilibrium expressions in equation (5) yields

$$E_c(\alpha) = \frac{(1-t)^2\left(\alpha^2(1+b-\alpha_0) + 2b\alpha_0 - 3b\alpha\alpha_0\right)}{2(b+\alpha)^2} \tag{6}$$

The FOC gives a unique solution

$$\tilde{\alpha}_c = \frac{\alpha_0(3b+4)}{\alpha_0+2b+2} > \alpha_0$$

and the SOC is verified.[11] If $\alpha_0 \geq 2/3$, $\tilde{\alpha}_c \geq 1$ and, since $E'(\alpha) < 0$ for all $\alpha \in (0,1]$, R's expected loss is minimized for $\alpha_c^\star = 1$ (as $\bar{\alpha} < 1$, the constraint $\alpha > \bar{\alpha}$ is then non-binding). When instead $\alpha_0 < 2/3$, R's expected loss is minimized for $\alpha_c^\star = \tilde{\alpha}_c$ provided $\tilde{\alpha}_c > \bar{\alpha}$, i.e. if $\alpha_0 > \frac{2-(b+1)2t}{t+3} = \frac{2}{3+t}\bar{\alpha}$, and for $\alpha_c^\star = \bar{\alpha}$, i.e. at the boundary, otherwise.

There are no confessors. When $\alpha \leq \bar{\alpha}$, replacing equilibrium expressions in equation (5) yields

$$E_{nc}(\alpha) = \frac{t\left(2\alpha_0(1-\alpha)^2 + (1+b)t\alpha^2 - t\alpha_0\left(2-(2-b-\alpha)\alpha\right)\right)}{2(1-\alpha)^2} \tag{7}$$

The FOC gives a unique solution

$$\tilde{\alpha}_{nc} = \frac{(2+b)\alpha_0}{2+2b-b\alpha_0} \in (0, \alpha_0)$$

and the SOC is verified.[12] Hence, $E_{nc}(\alpha)$ is minimized for $\alpha_{nc}^\star = \tilde{\alpha}_{nc}$ if $\tilde{\alpha}_{nc} < \bar{\alpha}$, i.e. if $\alpha_0 < \frac{2-(b+1)2t}{2-bt} = \frac{2}{2-bt}\bar{\alpha}$ and for $\alpha_{nc}^\star = \bar{\alpha}$, i.e. at the boundary, otherwise.

As $\alpha_c^\star > \alpha_0$ and $\alpha_{nc}^\star < \alpha_0$, R's global optimum α^* differs from α_0, which proves point i. If

[11] $$E_c''(\alpha)\big|_{\alpha=\tilde{\alpha}_c} = \frac{b(1-t)^2(\alpha_0+2b+2)^4}{16(b+1)^3(2\alpha_0+b)^3} > 0.$$

[12] $$E_{nc}''(\alpha)\big|_{\alpha=\tilde{\alpha}_{nc}} = \frac{t^2(2+b(2-\alpha_0))^4}{16(1+b)^3(1-\alpha_0)^3} > 0.$$

$\bar{\alpha} = 0$, $\alpha^* = \alpha_c^\star$. Suppose instead that $\bar{\alpha} > 0$. The previous considerations and the fact that R's expected loss is continuous in α with $E_c(\bar{\alpha}) = E_{nc}(\bar{\alpha})$ imply that whenever the minimum of a given case obtains at the boundary $\alpha = \bar{\alpha}$, the minimum of the other case is strictly lower. Indeed, if $\alpha_0 \leq \frac{2}{3+t}\bar{\alpha}$, $E_c(\alpha)$ is increasing in the whole $\alpha > \bar{\alpha}$ region and hence $\alpha^\star = \tilde{\alpha}_{nc}$, which proves point iii. Likewise, if $\alpha_0 \geq \frac{2}{2-bt}\bar{\alpha}$, $E_{nc}(\alpha)$ is increasing in the whole $\alpha \leq \bar{\alpha}$ and hence $\alpha^\star = \alpha_c^*$. Conversely, in the region $\alpha_0 \in \left(\frac{2}{3+t}\bar{\alpha}, \frac{2}{2-bt}\bar{\alpha}\right)$, α^* may be either $\alpha_c^\star$ or $\alpha_{nc}^\star$. Still, we now prove that $\alpha^* = \alpha_c^\star$ whenever $\alpha_0 > \bar{\alpha}$, so that point ii obtains.

Consider hence the case $\alpha_0 \in \left(\bar{\alpha}, \frac{2}{2-bt}\bar{\alpha}\right)$, or equivalently, $t \in (\underline{t}, \bar{t})$, where $\underline{t} \equiv \frac{1-\alpha_0}{b+1}$ and $\bar{t} \equiv \frac{2(1-\alpha_0)}{2+2b-b\alpha_0}$. Also, let $\Delta \equiv E_{nc}(\alpha_{nc}^\star) - E_c(\alpha_c^\star)$ be the difference in R's expected loss in the case without and with confessors given R's respective locally optimal choices, where $\alpha_{nc}^\star = \tilde{\alpha}_{nc}$ necessarily since $\alpha_0 < \frac{2}{2-bt}\bar{\alpha}$. When $\alpha_0 < 2/3$, using that $\alpha_c^\star = \tilde{\alpha}_c$,

$$\Delta = \frac{\alpha_0}{8(1+b)}\left(\frac{t(8(1+b)(1-t) - (8-4t+b(8-(4-b)t))\alpha_0)}{1-\alpha_0} - \frac{(1-t)^2(8+8b-8\alpha_0-9b\alpha_0)}{b+2\alpha_0}\right).$$

The expression is strictly positive, so that $\alpha^* = \alpha_c^\star$, since it is concave in t and strictly positive in the two extrema (the symbol $\propto$ means "has the same sign as"):

$$\Delta\big|_{t=\underline{t}} \propto \alpha_0^2 b\left(4b+4-2\alpha_0-3\alpha_0 b\right)^2 > 0,$$
$$\Delta\big|_{t=\bar{t}} \propto \left(2-\alpha_0\right)^2 + \alpha_0 b^2 + 2\left(2-\left(2-\alpha_0\right)\alpha_0\right) b > 0.$$

When $\alpha_0 \geq 2/3$, instead, using that $\alpha_c^\star = 1$,

$$\Delta = \frac{8t(1-\alpha_0)(1+b\alpha_0) - (t^2(4+b\alpha_0(8-(4-b)\alpha_0)) + 4(1-\alpha_0)^2)}{8(1+b)(1-\alpha_0)}.$$

Again, the expression is strictly positive, so that $\alpha^* = \alpha_c^\star$, as it is concave in t and strictly positive at the two extrema. Indeed, $\Delta\big|_{t=\underline{t}} \propto 4\alpha_0^2 - b\alpha_0^2 + 8b\alpha_0 - 4b$ which is increasing in α_0 and equal to $\frac{16+8b}{9} > 0$ in $\alpha_0 = 2/3$. Likewise, $\Delta\big|_{t=\bar{t}} \propto 2\alpha_0^2 + b\left(3\left(2-\alpha_0\right)\alpha_0 - 2\right)$ which is increasing in α_0 and equal to $\frac{2b}{3} + \frac{8}{9} > 0$ in $\alpha_0 = 2/3$.

A.2 Proof of proposition 3

Simple calculation shows that the optimal cutoff $z^*(\alpha, t, b)$ (for the exact value see the end of section 4.1.1) is always larger than the equilibrium value of $\bar{y}(\alpha, t, b)$ without delegation. It follows that R never wants to overrule the decision of α_n when α_n decides to exonerate S knowing only that there was no confession and that $z > z^*(\alpha, t, b)$. Clearly, R does not want to overrule when after a confession α_n decides to accuse S. In fact, R could even learn the value

of z when it is above $z^*(\alpha, t, b)$. However, this is not always the case when communicating with α_τ. Clearly, R does not want to overrule α_τ when α_τ decides to exonerate S. To check that R does not want to overrule α_τ when α_τ decides to accuse S one must find the largest $\tilde{z}$ for which

$$\alpha t(\tilde{z} - t) \geq \frac{1}{2}(1 - \alpha)(\tilde{z} - t)(\frac{\tilde{z} + bt}{1 + b} - t)$$

and show that $z^*(\alpha, t, b)$ is smaller. Simple calculation shows that indeed this is the case even if $\bar{t} = \frac{1-\alpha}{1+b}$ or if $\bar{t} = \frac{1-\alpha}{(1+\alpha)(1+b)}$ (for the definition of $\bar{t}$ see again the end of section 4.1.1). Note that the multiplier $1/2$ of the l.h.s. would disappear (and also $(\tilde{z} - t)$ from both sides) if R knew the value of z when it is equal to the smallest innocent type who separates.

A.3 Alternative delegation policies

A.3.1 Random delegation

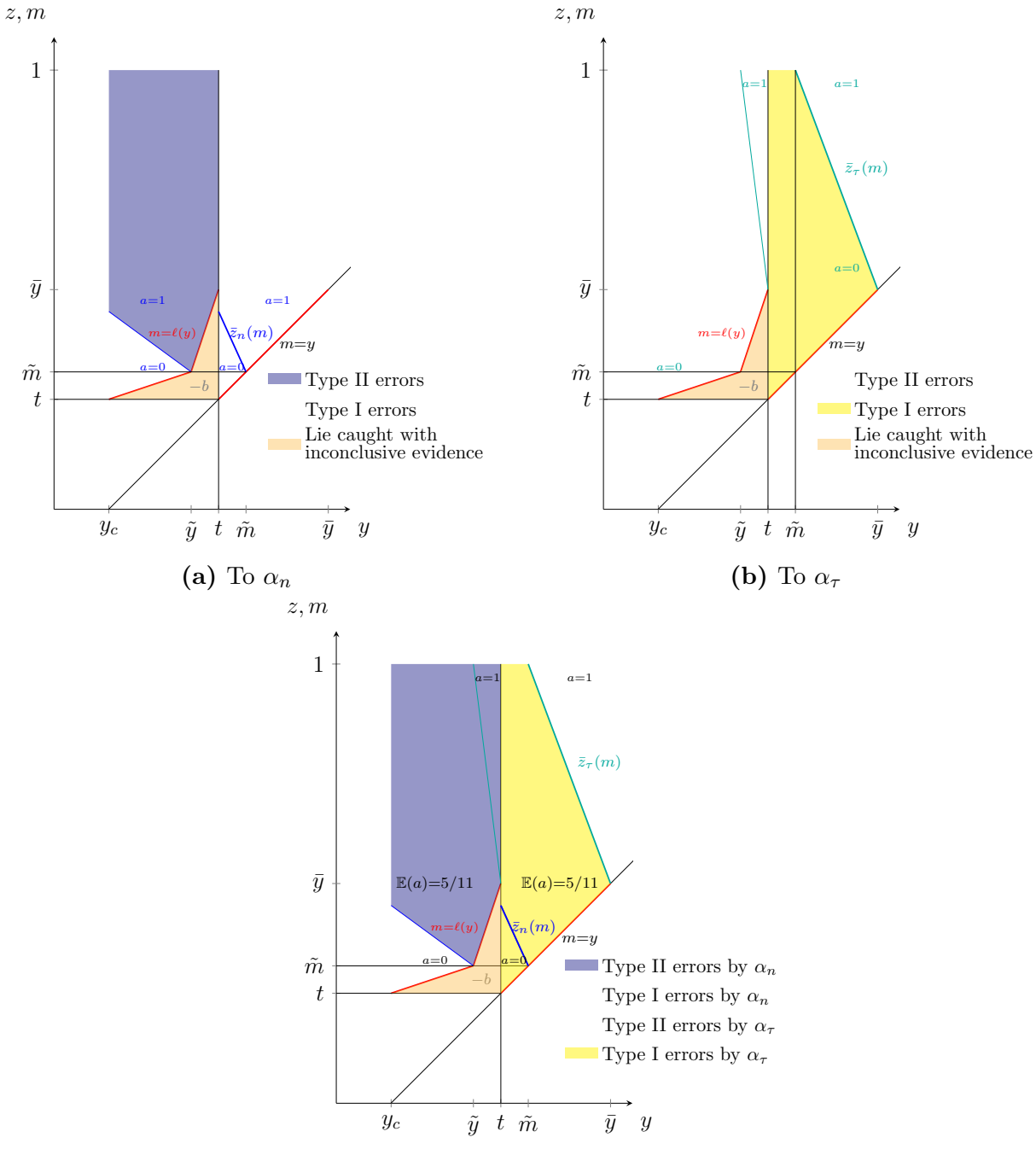
Let us first consider the following conditional delegation policy

$$\alpha(m, z) = \begin{cases} \alpha_n & \text{if } z \geq z(m) \\ \alpha_\tau & \text{if } z < z(m), \end{cases}$$

where $\mathbf{z} : [0, 1] \to [0, 1]$ is a linear cut-off function given by $z(\tilde{m}) = \tilde{z}$ and $z(1) = 1$. Hence, the probability that the nice interrogator α_n makes the decision conditional on S is not caught in a lie is constant for each m and equal to $\frac{1-\tilde{z}}{1-\tilde{m}}$. Given that S only cares about the distribution of types of interrogators given some message m, it follows that the delegation decision can be made at once at the very beginning of the interrogation even without knowledge of the evidence or of the denying message. In our example, the interrogation should be given to the nice interrogator with probability 5/11. Naturally, S must again be kept in the dark in that he should not observe the preferences of the actual interrogator. The equilibrium cut-off strategies $\bar{\boldsymbol{z}}_{\boldsymbol{n}}$ and $\bar{\boldsymbol{z}}_{\boldsymbol{\tau}}$ of the nice and tough interrogator, respectively, and the resulting type I and type II errors made by α_n and α_τ are displayed in figure 3.

A.3.2 Reverse delegation

When α_n and α_τ are assigned to strong and weak evidence, respectively, one can achieve the same payoffs as in the opposite case. In this case, however, the cut-off value changes to $\tilde{z}_r$, and α_n and α_τ follow also cut-off strategies described by $\bar{z}_r(m)$, with the difference that they choose action 1 if $z \leq \bar{z}_r(m)$ and action 0 otherwise. We illustrate this new equilibrium in figure 4 and



(a) To α_n

(b) To α_τ

(c) Superimposing 3a and 3b with probabilities 5/11 and 6/11

Figure 3 Random delegation

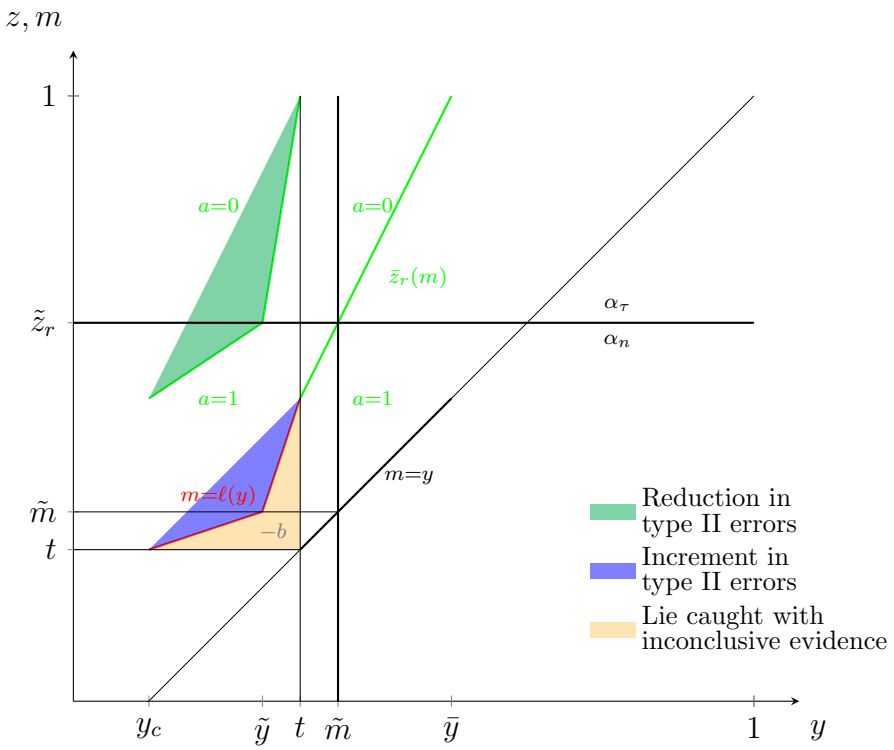


Figure 4 Reverse delegation

the total reduction in type II errors as the difference of the green and the blue area, which is just equal to the green area in figure 2.

A.4 Proof of proposition 4

We first describe players' equilibrium strategies and then verify sequential rationality along the equilibrium path (since beliefs are free off the path, we can always make sure that decisions specified there are sequentially rational). Throughout, all specified beliefs are consistent with the generalized version of Bayes rule, and $g(m) \equiv \ell^{-1}(m)$.

Candidate equilibrium strategies. Let $\bar{y}$, $\bar{z}(m)$ and S's behavioral strategy in stage 1 be as in proposition 0. R always chooses $a = 0$ if $m < t$ and $a = 1$ if $m \geq \bar{y}$ (provided S is not caught in a lie, otherwise off the equilibrium path R again chooses $a = 0$ and S gets $-b$). Instead, for $m \in [t, \bar{y})$, R chooses $a = 1$ if $z \geq Z(m)$ and delegate to α_τ if $z < Z(m)$ where

$$\begin{aligned} Z(m) &= \bar{z}(m) + b(m - g(m)) = \bar{z}\left(g\left(m\right)\right) \ \in (\bar{z}(m), 1) \\ &= \bar{y} - b(t - \bar{y} + y_c) + \frac{bt}{\alpha} - \frac{b(1-\alpha)}{\alpha} m. \end{aligned} \tag{8}$$

Consider now stage 3 after message m was sent and R delegated in accordance with the strategy above. Then, α_τ chooses $\mathcal{M}_\tau = \{g(m), m\}$ and S sends $m_2 = g(m)$ if guilty and $m_2 = m$ if

innocent. α_τ recommends $a_2 = 0$ if $m_2 = g(m)$ (and off the equilibrium path informs R if S is caught in a lie, in which case S will get $-b$ as in the baseline model). If $m_2 = m$ then α_τ chooses the recommendation a_2 according to $\bar{z}(m)$. Finally, assume that if R delegates when she should not given $Z(m)$, α_τ always chooses $a_2 = 0$, so that this recommendation is uninformative about the second message of S.

Sequential rationality. R's strategy upon a pooling message is sequentially rational:

- if S is caught in a lie, R believes that S is surely guilty and anticipates he will confess honestly to α_τ who will recommend $a_2 = 0$;
- if S is not caught in a lie, R believes S is innocent with probability α and:
  - when $z \geq Z(m)$, she is hence indifferent to any action or delegate to α_τ, who will recommend $a_2 = 0$;
  - when $\bar{z}(m) \leq z < Z(m)$, she strictly prefers to delegate since she will make no error at all since α_τ will recommend $a_2 = 0$ if S is guilty and $a_2 = 1$ if S is innocent;
  - when $z < \bar{z}(m) < Z(m)$, R knows that α_τ will recommend $a_2 = 0$, no matter if S is guilty or innocent. Given that R believes S is innocent with probability α she is again just indifferent between delegating and choosing $a = 1$.

α_τ's strategy is also sequentially rational together with the belief that S is surely innocent in the only instance in which she does not recommend $a_2 = 0$ and knowing that R will follow the recommendation.

Finally, consider S's strategy. When interrogated by α_τ, the strategy of innocent type m is clearly optimal. As for a guilty type $g(m)$, given his belief that $z < Z(m)$, by construction he is now indifferent between confessing honestly, which yields 0, and sending $m_2 = m$, since his expected payoff from doing so is $-b\left(m - g\left(m\right)\right) + Z(m) - \bar{z}\left(m\right) = 0$. The optimal mechanism $\hat{\boldsymbol{z}}^\star$ is given by $1 - \hat{z}^\star(y) = 1 - \bar{z}(m(y)) - \left(m\left(y\right) - y\right) b = 1 - \bar{z}(y)$ as calculated in Ispano and Vida (2021). Besides, guilty types and innocent types who separate in the equilibrium of the baseline model still get always 0 and 1, respectively, i.e. $\hat{z}^\star(y) = 1$ for $y < y_c$ and $\hat{z}^\star(y) = y$ for $y \geq \bar{y}$. Consider now stage 1 and notice that for each type y the joint on the equilibrium path behavior of R and α_τ is in expectation equivalent to the optimal mechanism $\hat{\boldsymbol{z}}^\star$. In particular, for pooling innocent types $\bar{z}(y) = \hat{z}^\star(y)$ and for pooling guilty types $\bar{z}(g(m)) = \bar{z}(y) = \hat{z}^\star(y)$. It follows that no type y can benefit from playing as if he was some other type y'' throughout the game otherwise she would do so in the optimal mechanism as well. Finally, no type can

profit from deviating at stage 1 to some pooling message m' and then send $m_2 = m'$ in stage 2. Indeed the choice of $Z(m)$ is such that it is as if this type was deviating in the equilibrium of the baseline model, where it is also the case that $a(m, z) = 1$ whenever $z \geq Z(m)$. In short, S can either behave as if he was another type or lie and stick to his stage 1 story. In the first case, it is as if he was playing in the optimal mechanism, hence this type of deviation is not profitable. In the second case, it is exactly as he was playing in the equilibrium of the baseline model, so that this type of deviation is again not profitable. Finally, R always finds it optimal to follow the recommendation of α_τ. This is clear when exoneration is recommended or when R knows that she will not make a mistake. When accusation is recommended and R is uncertain about S's type, her belief is α so indeed she will follow the recommendation.