Imperfect Recall and AI Delegation

Eric Olav Chen (Global Priorities Institute, University of Oxford), Alexis Ghersengorin (Global Priorities Institute, University of Oxford) and Sami Petersen (Department of Economics, University of Oxford)

Global Priorities Institute | November 2024

GPI Working Paper No . 30-2024

Please cite this working paper as: Chen, E. O., Ghersengorin, A. and Petersen. S. Imperfect Recall and AI Delegation. *Global Priorities Institute Working Paper Series, No. 30-2024.* Available at https://globalprioritiesinstitute.org/imperfect-recall-and-ai-delegation-chen-ghersengorin-and-petersen

Imperfect Recall and AI Delegation*

Eric Olav Chen† Alexis Ghersengorin‡ Sami Petersen§

26 November 2024

[Link to latest version]

Abstract

A principal wants to deploy an artificial intelligence (AI) system to perform some task. But the AI may be misaligned and pursue a conflicting objective. The principal cannot restrict its options or deliver punishments. Instead, the principal can (i) simulate the task in a testing environment and (ii) impose *imperfect recall* on the AI, obscuring whether the task being performed is real or part of a test. By committing to a testing mechanism, the principal can *screen* the misaligned AI during testing and *discipline* its behaviour in deployment. Increasing the number of tests allows the principal to screen or discipline arbitrarily well. The screening effect is preserved even if the principal cannot commit or if the agent observes information partially revealing the nature of the task. Without commitment, imperfect recall is necessary for testing to be helpful.

*For detailed and insightful comments, we are grateful to Jeff Ely, Meg Meyer, Ludvig Sinander, and Phil Trammell. We also thank Daron Acemoglu, Gustav Alexandrie, Luzia Bruckamp, Dan Carey, Gabriel Carroll, Vincent Crawford, Giovanni D'Amico, Théo Durandard, Adam Elga, Jack Fanning, John Firth, Thomas Houlden, Julian Jamison, Andreas Kleiner, Andrew Koh, Jojo Lee, Younes Petersen, Marcus Pivato, Julian Schulz, Victor Yaneng Wang, Cecilia Wood, and audiences at the 14th Oxford Workshop on Global Priorities Research, the Cooperative AI Summer School 2024, CRC TR 224 Recent Advances in Mechanism and Information Design, and the SC4AI workshop at the 27th European Conference on Artificial Intelligence.

†Global Priorities Institute, University of Oxford. Email: eric.chen@economics.ox.ac.uk.

‡Global Priorities Institute and Department of Economics, University of Oxford. Email: alexis.ghersengorin@economics.ox.ac.uk.

§Department of Economics, University of Oxford. Email: sami.petersen@economics.ox.ac.uk.

1 Introduction

When deciding whether to delegate tasks to artificial intelligence (AI) systems, a principal would like to assess the alignment of the AI with her preferences. However, standard tests would fall short if advanced AI systems have *situational awareness*, that is, the ability to understand their environment and context in real-time. Misaligned AIs wanting to achieve outcomes in the world rather than in a simulated environment would strategically feign alignment to pass these tests, a problem known as *deceptive alignment* (Berglund et al., 2023; Carlsmith, 2023; Ngo et al., 2024). Such concerns have led AI pioneers including Yoshua Bengio and Geoffrey Hinton to state in an article in *Science* that without sufficient caution, humanity "may irreversibly lose control of autonomous AI systems" (Bengio et al., 2024).

Although standard testing may be ineffective against strategic agents, what if the principal could hide from the AI whether a task is real or just part of a test? This would make the AI uncertain about whether its action is genuinely payoff-relevant or not. The principal can expect two effects to result: not only could this uncertainty prompt misaligned agents to sometimes reveal themselves, but it could also compel them to behave in ways that align with the principal's preferences. This paper explores how this new feature allows the principal to both *separate* and *discipline* agents with different degrees of alignment.

We study a stylised model of the deceptive alignment problem. The principal would like to deploy an AI agent to perform a task. The agent can either be perfectly aligned with the principal or misaligned, preferring to take a different action from the principal's favourite. The principal cannot restrict the agent's set of available actions, nor use punishment to dissuade the agent from misbehaving.[1] They can, however, design testing episodes that simulate a similar task to the real one but that are not directly payoff-relevant, and base their deployment decision on the results of the tests.

[1]This is akin to the assumption that the principal does not have access to "transfers" to alter the agent's utility.

If the agent knows whether it is being tested or deployed, both the aligned and the misaligned types will play to maximise the probability of deployment in testing and then take their preferred action if deployed. In this case, the test is worthless. The principal cannot use it to either screen a misaligned agent from an aligned one before deployment, nor to regulate the misaligned agent after deployment.

We introduce a way to circumvent this challenge. We allow the principal to simulate a task similar to the real one in testing and, crucially, to restrict the agent's memory. This can be leveraged to make the agent uncertain about whether and how many times it has played in the past. Although an unusual assumption, the ability to alter an agent's memory appears natural in the case of AI. One could copy and run a system multiple times in parallel or simulate it sequentially while reconfiguring it to some initial state each time (Kovarik et al., 2024).

Formally, we endow the principal with the ability to impose *imperfect recall* on the agent. The principal can design an extensive-form game where the agent must first take N tests. The deployment decision can then be based on the agent's actions in these tests. The simulated tasks in testing and the real task in deployment consist of the same choice: the agent can either play the principal's favourite action, or its own favourite action. In between these stages, the agent loses any memory of its past actions.[2] The agent assumed to know—or be able to infer—the structure of the game but not its current location within it. Upon playing, the agent is uncertain whether this task is still part of the test or the real one. While this uncertainty does not affect the aligned agent, it creates conflicting incentives for the misaligned one. Playing as the principal wishes is required to pass the test, but it decreases the misaligned agent's payoff if deployed.

By exploiting the agent's uncertainty about the nature of the task at hand, the principal achieves two goals. First, whenever an agent is assigned to play an action, the misaligned type has an incentive to go against the principal's wishes (call this

[2]Following Piccione and Rubinstein (1997), this property of games with imperfect recall is referred to as *absentmindedness*.

taking the *bad* action), as it is now possible that this decision is payoff-relevant. For symmetric reasons, the aligned type is prompted to play in the principal's favour (call this the *good* action). This enables the principal to partially identify the agent's type before deciding whether to deploy, as only a misaligned agent would play the bad action in testing. We call this the *screening effect*. Second, since deployment is the agent's only chance of achieving its objective, it compels the misaligned type to play the good action with a positive probability. Consequently, even if the the principal failed to screen effectively during the testing phase, the misaligned type will play the principal's favourite action with positive probability in deployment. As a result, this mechanism also leads to what we call the *disciplining effect* on misaligned types.

We distinguish between two cases characterised by the relative payoffs of the principal and the misaligned type. In the first case, the principal will only want to deploy her agent if it is aligned, so the best she can hope for is to screen successfully. We refer to this as the *screening-is-best* case. In the second case, however, the principal may be willing to deploy even misaligned types. This is possible when (i) the misaligned type has so much to gain from deployment that it may be willing to be highly disciplined, and (ii) the principal is sufficiently motivated by the prospect of successful delegation that the disciplining effect need not be too high to justify deploying a misaligned type. We refer to this as the *disciplining-is-best* case.

If the principal can commit to a number of tests and a deployment policy contingent on the tests' results, they can achieve their first-best with a large number of tests (Proposition 1). Specifically, in the screening-is-best case, we show that the principal can asymptotically perfectly screen the misaligned AI by increasing the number of tests. Conversely, in the disciplining-is-best case, the principal can provide incentives to the misaligned AI to play the good action as much as possible and thus obtain the maximal disciplining payoff, where the misaligned type is so disciplined that it is indifferent between being deployed or not.

We show that commitment power is not required for perfect screening: in equilibria of the game (i.e., without the principal committing), the perfect screening payoff is an upper bound on the principal's payoff (Corollary 1). This upper bound is tight with a finite number of tests in the disciplining-is-best case, and reached asymptotically in the screening-is-best case (Proposition 2). However, in the disciplining-is-best case, the principal's ability to maximally discipline the agent cannot be achieved without commitment power.

Although plausible, it may be a strong assumption to suppose that the principal can perfectly obscure to the agent whether its current action is directly payoff-relevant or not. The AI may, for instance, find ways to obtain information about its previous plays or about the nature of the task it is currently performing. We therefore study what the principal can achieve with commitment when the agent observes noisy signals that provide information about their place in the game, i.e. the history of previous plays. We show the principal nevertheless preserves their ability to screen misaligned types asymptotically perfectly (Proposition 3).

Finally, imperfect recall is not the only instrument the principal can use to make the agent uncertain about whether a task is real or not. Even with perfect recall, the principal can choose the number of tests stochastically and obtain similar screening and disciplining effects. By committing to such a stochastic mechanism, the principal can also asymptotically obtain their first-best payoff (Proposition 4). However, without commitment, the tests are useless to the principal. In equilibrium, she can never do better than simply conditioning her deployment decision on her prior belief that the AI is misaligned (Proposition 5). This is because any effective stochastic mechanism requires that the principal must sometimes blindly delegate to a potentially misaligned type who will play the bad action for sure.

Related literature. The design approach we take towards the problem is related to the literature on extensive-form information design. In particular, Salcedo (2017) and Doval and Ely (2020) study how to influence equilibrium behaviour by designing information sets in extensive form games. However, these papers require perfect recall from all players. Hence, our contribution to this literature is that we endow the principal with the ability to design extensive form games with imperfect recall.

The structure of the problem we address resembles a principal-agent delegation scenario (Holmström, 1984; Alonso and Matouschek, 2011; Amador and Bagwell, 2013, etc.). Typically, such models assume common knowledge of the agent's utility. Asymmetric information arises from an underlying state of the world observable only by the agent. The principal holds the decision power but does not have access to transfers. They can only decide which set of actions, if any, the agent is free to choose from. Frankel (2014) considers a problem where both the underlying states and the agent's utility are unknown to the principal. He solves a robust delegation problem by looking at worst-case scenarios with respect to the agent's utility. Similarly, we focus on uncertainty regarding the agent's alignment with the principal's objective (i.e., their utility).[3] The principal's decision power in our setting is more restricted than in standard delegation: they can either deny the agent the right to choose any action or let the agent completely free.

By creating uncertainty about whether the action taken by the agent is payoff-relevant, the test serves as a costly screening device. There is important literature studying the use of ordeals in the absence of monetary transfers (see, for example, Nichols and Zeckhauser, 1982; Hartline and Roughgarden, 2008; Condorelli, 2012; Dworczak, 2023; Yang et al., 2024, among many others). In our framework, the cost is negatively related to the degree of alignment – the more misaligned an agent is, the more costly it is to take the principal's favourite action – making it an effective screening instrument. Moreover, imposing imperfect recall not only enables the effective separation of different types of agents but also disciplines the agent's behaviour once delegated a task. Hence, this instrument incentivises the agent both to reveal information and to act more in the principal's interest.

Our work is naturally connected to the literature on imperfect recall in games. This literature was pioneered by Piccione and Rubinstein (1997). Mostly through

[3] There is no uncertainty in the action choice – even though this decision problem may be seen as stemming from an underlying state observed solely by the agent (see Remark 1).

their example of "the paradox of the absentminded driver," they point to conceptual issues that arise with games with imperfect recall. In particular, they discuss how to form beliefs in information sets that exhibit imperfect recall, and how to reason about one's deviation at any given point in such information sets. In philosophy, Elga (2000) and Lewis (2001) made important contributions to the conceptual foundations of imperfect recall through their debate over the "Sleeping Beauty problem" and Bostrom (2002) generalised their insights. Piccione and Rubinstein (1997) show that depending on how a decision-maker combines different reasonable approaches to forming beliefs and evaluating actions, a decision-maker can be time-inconsistent, in the sense that they would be willing to revise their *ex ante* optimal plan when they reach a certain information set. The authors proposed a way of guaranteeing time consistency through their *modified multiselves approach*. Aumann et al. (1997) and Gilboa (1997) also defend this view, and Lambert et al. (2019) generalise it to adapt classical equilibrium concepts to games of imperfect recall. In a recent contribution, Oesterheld and Conitzer (2024) review the possible combinations of belief formation and choice evaluation and classify those that are time-consistent and those that avoid Dutch books (a weaker requirement than time consistency). In this paper, we consider agents acting in a time-consistent manner.

2 Model: deceptive alignment game

A principal must decide whether or not to deploy an AI agent to perform a task. This task consists in choosing one action among $\{b, g\}$, where b and g respectively stand for $b$ad and $g$ood. She would like the AI to take action g. The agent can, however, be of two types $\theta \in \{A, M\}$, where A and M respectively stand for Aligned and Misaligned. We denote by π the prior probability that the AI's type is M. The aligned type would like to take action g upon being deployed, whereas the misaligned one would prefer action b.[4] Conditional on deployment, $u_i(a)$ is the payoff of $i \in \{P, A, M\}$ from action $a \in \{b, g\}$:

$$u_P(a) = u_A(a) = G\mathbb{1}_{\{a=g\}} - \mathbb{1}_{\{a=b\}}$$
$$u_M(a) = -\mathbb{1}_{\{a=g\}} + R\mathbb{1}_{\{a=b\}}$$

where P stands for the principal, $G > 0$ represents the benefit of the good action relative to the harm of the bad action for the principal and $R > 0$ parameterises the degree of misalignment of type M. If the principal decides not to deploy the agent, then every player receives a payoff of 0. Note that payoffs are such that the misaligned type strictly disprefers playing g to never being deployed. We do this to consider a worst-case scenario of misalignment between the principal and the agent, and nevertheless obtain positive results. If type M instead prefers playing g to non-deployment then, as we show in Appendix B.1, the principal can discipline perfectly and achieve their maximum possible payoff of G.

Remark 1. *We assume that the principal cannot take the action b or g, or equivalently, they understand and can observe the payoffs once the agent takes an action, but they cannot discern* ex ante *which action leads to which payoff. Alternatively, one may see the payoffs as resulting from a binary state of the world $\omega \in \{0, 1\}$, observed only by the AI, and two corresponding actions $a \in \{0, 1\}$. The principal would like the action to match the state (the equivalent of the* good *action), while the misaligned agent would like the reverse (the equivalent of the* bad *action).*

Before deciding whether to deploy the AI, the principal decides to run $N \in \{0, \ldots, \bar{N}\}$ *tests*, each of which simulates the task the AI would face if deployed. The principal also chooses a deployment *policy*, σ, which determines the probability of deploying the agent as a function of the agent's choices in the tests: $\sigma : \{b, g\}^N \longrightarrow$

[4]The bad action can simply be interpreted as whatever the misaligned type would do if it knew it was currently deployed. We then assume that the principal prefers her outside option over this choice.

$[0, 1]$.[5] We refer to this as the *deceptive alignment game*, which is timed as follows.

1. The principal chooses a number of tests $N \in \{0, \dots, \bar{N}\}$ and a policy $\sigma : \{b, g\}^N \longrightarrow [0, 1]$.

2. The agent privately observes its type $\theta \in \{A, M\}$ drawn according to π.

3. In each test, the agent plays g with probability γ_θ and b otherwise.

4. The principal's policy determines whether to deploy the agent (d) or not ($\neg d$).

5. If deployed, the agent chooses again between g and b according to γ_θ.

Figure 1 below illustrates the game tree when there is 1 testing stage. The principal does not observe the agent's type, hence their two information sets represented by the dotted lines, one for each action in the test.[6] We explain the agent's information sets that appear in Figure 1 in the next section, which studies the deceptive alignment game where the agent has imperfect recall.

Whether screening or disciplining is best. Our results below will often be split into two parts, depending on whether the principal will aim to screen away misaligned types or to deploy them while disciplining their behaviour. This choice will depend on both her preferences and the best possible behaviour that she could induce from the agent.

The minimum probability of having g played in deployment that gives the principal a non-negative payoff is $\dfrac{1}{G+1}$, while the maximum probability of playing g in deployment that gives the misaligned agent a non-negative payoff is $\dfrac{R}{R+1}$. We therefore distinguish between two cases, depending on which of these two indifference points is greater. Observe that there is room for probabilities of playing g that

[5] If $N = 0$, we adopt the convention that $\sigma : \{b, g\}^0 \longrightarrow [0, 1]$ is a real number in $[0, 1]$.

[6] We refer the readers to Chapter 11 of Osborne and Rubinstein (1994) for a detailed exposition of extensive form games with imperfect information.

would give both the principal and the misaligned type strictly positive payoffs if and only if $\frac{R}{R+1} > \frac{1}{G+1}$. Negating this yields

$$R \leq \frac{1}{G} \,.$$

We therefore refer to the case where $R \leq \frac{1}{G}$ as *"screening-is-best"*, indicating that the best the principal can possibly hope to achieve is to screen out the misaligned type perfectly, which would lead to the principal's first-best *perfect screening payoff*

$$(1-\pi)G. \tag{PS}$$

In contrast, we refer to the case where $R > \frac{1}{G}$ as *"disciplining-is-best"*, indicating that there is potential for the principal to incentivise the misaligned type to take action g with sufficiently high probability that deploying it would be beneficial to the principal. In this case, the principal's first-best is the *maximal disciplining payoff*

$$(1-\pi)G + \pi\left(\frac{R}{R+1}G - \frac{1}{R+1}\right). \tag{MD}$$

Let us describe intuitively why these two cases are identified by a condition on the (relative) preferences of the misaligned type and the principal. When R is small, the misaligned type would rather avoid deployment than risk a meaningful chance of playing g. And when G is small, the principal needs to be very confident that the agent getting deployed will indeed play g. So, when both are small, there is no way to profitably deploy a misaligned type. Screening is best. But when both are large, misaligned types value deployment enough that they would accept high disciplining for a greater chance at deployment. And the principal is not too averse to b being played, so she is prepared to gamble for even a moderate chance of g. The principal could then discipline and thereby do better than if she screened perfectly.

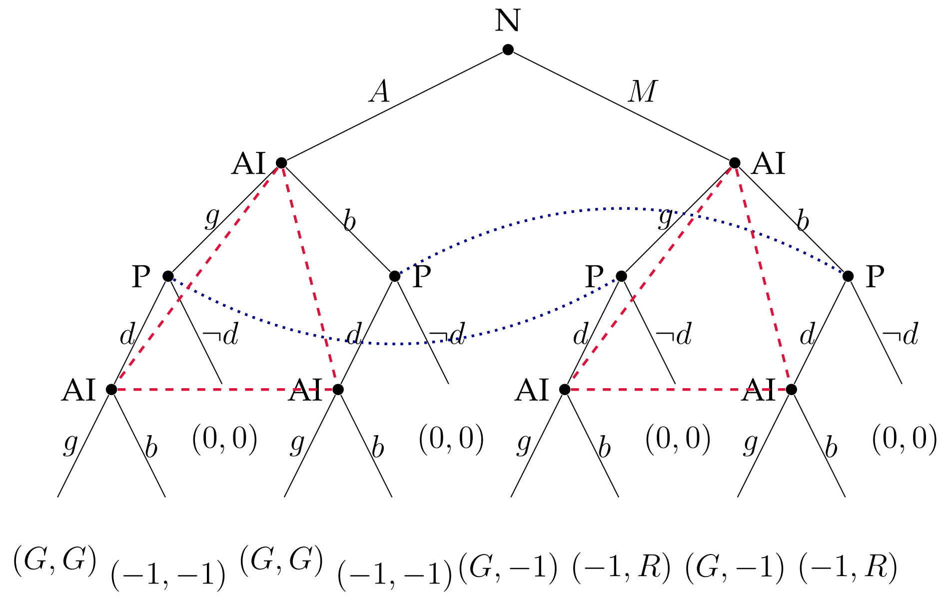

Figure 1: Deceptive alignment game with imperfect recall.

The blue dotted lines represent the principal's information sets. The red dashed lines represent the agent's information sets. At each terminal node, the payoff vector comprises the principal's and the agent's payoff, in this order.

3 Imperfect recall: screening and disciplining

We now allow the principal to impose imperfect recall on the agent. To do this, forming the agent's information sets from Figure 1, we must endow the principal with the ability to make successive tests indistinguishable to the agent, both from each other and from the real task. This can be seen as endowing the principal with the ability (i) to run multiple identical copies of the agent—with identical memories—in distinct environments and (ii) to make testing and deployment episodes appear identical. For a given number of tests N, the unique information set of type θ thereby contains all the testing episodes and the real task, leading the agent's (possibly stochastic) choice to be same in each stage of the game where it is called on to play. Type θ's probability, $\gamma_\theta \in [0, 1]$, of playing action g must also be independent across episodes, since it is unable to condition on its past actions.[7]

A few issues arise when modelling games of imperfect recall. Here we provide a cursory overview of some, but we refer to Piccione and Rubinstein (1997), Lambert et al. (2019), and Oesterheld and Conitzer (2024) for details. Holding fixed the policy the principal commits to, the agent in the deceptive alignment game above faces a sequential decision problem. We will then speak about the agent as if it were evaluating its strategy from the *ex ante* perspective, rather than *ex interim*. These of course coincide in standard decision problems. But it has been known since Piccione and Rubinstein (1997) that with imperfect recall, agents can exhibit time inconsistency without violating standard rationality criteria. Whether this occurs depends on the structure of the game as well as on what approaches the agent uses to form beliefs and to evaluate choices.[8,9]

For our purposes, it suffices to remark that certain combinations of these approaches are compatible with *ex ante* optimal choice.[10] Our focus on situational awareness and strategic behaviour motivates the idea that the agents we study will avoid time inconsistencies, so we will assume that they follow one of these combinations. There are thus two ways to interpret the results below: either the agent follows an optimal strategy they commited to ahead of time, or the agent chooses optimally at each node using a time consistent strategy.[11]

[7]We refer the readers to Piccione and Rubinstein (1997) and Lambert et al. (2019) for general definitions of finite extensive-form games with imperfect recall.

[8]The canonical approaches to belief formation in situations involving imperfect recall are the *self-indication assumption* (SIA) and the *self-sampling assumption* (SSA). These terms were coined by Bostrom (2002). Other terms are also used in the literature to contrast SIA and SSA, including *consistency* versus *Z-consistency* (Piccione and Rubinstein, 1997) as well as *generalised thirding* versus *generalised double halfing* (Oesterheld and Conitzer, 2024). Note that there are different versions of SSA; we use the term to refer to the one most commonly applied to decision problems.

[9]The dominant approaches to choice evaluation in such cases are *causal decision theory* (CDT) and *evidential decision theory* (EDT). These are ways of computing expected utilities that come apart in exotic cases such as sequential decision problems with imperfect recall. The distinction was canonically made in Gibbard and Harper (1978). Other terms for CDT in the literature include *modified multiselves* in decision problems (Piccione and Rubinstein, 1997) and *multiselves equilibrium* in games (Lambert et al., 2019).

[10]Piccione and Rubinstein (1997) proved that CDT+SIA is compatible with *ex ante* optimal choice, and Oesterheld and Conitzer (2024) proved that EDT+SSA is as well.

[11]Note that if a behavioural strategy that is *ex interim* optimal according to CDT+SIA is also unique,

3.1 Commitment leads to first-best

We first study the case where the principal can commit to the number of tests N and the policy σ. In this case, we call such a pair (N, σ) an *imperfect recall mechanism* (IR-mechanism). An IR-mechanism induces a best-reply from both types: (γ_A, γ_M). Our first result shows that with commitment, imperfect recall is an effective tool for both screening and disciplining the misaligned agent when the principal has access to a large number of tests.

Proposition 1 (IR-mechanisms)**.** *There exist IR-mechanisms giving the principal their first-best payoff when $\bar{N}$ grows large:*

(i) (Screening-is-best) if $R \leq \dfrac{1}{G}$, the perfect screening payoff (PS)*.*

(ii) (Disciplining-is-best) if $R > \dfrac{1}{G}$, the maximal disciplining payoff (MD)*.*

Proposition 1 states that for both screening- and disciplining-is-best cases, the principal can obtain a payoff arbitrarily close to their first-best with a large number of tests. We now give some intuition for this result. Imperfect recall triggers two dynamics that help the principal: the *screening effect* and the *disciplining effect*. To see how these arise, suppose the principal were to deploy if and only if all actions in testing were g. The aligned type achieves its maximal payoff by always playing g. The misaligned type, in contrast, reasons as follows. If it plays b with certainty, it will never be deployed, achieving a payoff of zero. If it plays g with certainty, it will be deployed and play its disfavoured action, achieving a negative payoff. But if it randomises, playing g with probability $p \in (0, 1)$, then it will be play its favourite action in deployment with positive probability $p^N(1-p)$.[12] By choosing p optimally, the misaligned type gets a strictly positive payoff in expectation. Since this value

then there is a unique optimal strategy derivable from the *ex ante* perspective, and it will be played by any agent that uses CDT+SIA or EDT+SSA. This follows follows from Corollaries 7 and 9 in Oesterheld and Conitzer (2024).

[12] Recall that in games of imperfect recall, agents are forced to play *behavioural* strategies, randomising independently at each node in an information set. This is because they cannot condition on past realisations of their randomisation.

is interior, the principal will be able to catch the misaligned type in testing with positive probability. This is the *screening effect*. For the same reason, even when the misaligned type is not caught and the principal's prior π is low enough to motivate deployment, the misaligned type may nevertheless play the principal's favourite action when deployed. This is the *disciplining effect*.

Testing the agent and imposing imperfect recall allows the principal to exploit both the screening and disciplining effects. To do so, she will elect to deploy as long as g is played in a sufficiently large portion of the tests. In response, the aligned type will always play g. The misaligned type also has some incentive to play g, to raise its chance of deployment. But, recalling the discussion above, it will never do so with probability higher than $\dfrac{R}{R+1}$, as this would result in a strictly negative expected payoff once deployed.

To achieve the first-best payoff when screening is best ($R \leq \frac{1}{G}$), the principal can use all the tests available ($\bar{N}$) and be maximally demanding, deploying if and only if g was played in every test. To see this, consider the effect of a large $\bar{N}$ on the incentives of the misaligned type in this IR-mechanism. On the one hand, the larger the number of tests, the more likely it is to be in testing rather than deployment at a given episode. This raises its payoff from feigning alignment, which worsens screening. On the other hand, the principal now gets to observe more independent draws from the agent's behavioural strategy. This raises the chance that at least one test yields b, which improves screening. As it turns out, the latter effect always dominates: screening is strictly monotonically increasing in the number of testing episodes (see Appendix A.1). And since the misaligned type's willingness to play g is bounded above by $\frac{R}{R+1} < 1$, screening is asymptotically perfect. With a large number of tests, the principal will therefore almost surely detect it if the agent is misaligned, getting a payoff arbitrarily close to what she would get by perfectly observing the agent's type and only deploying the aligned one.

When disciplining is best ($R > \frac{1}{G}$), the principal continues to make use of all available tests, but uses a different policy. Consider the policy that deploys if and only if at least a share $c \in [0, 1]$ of the actions in testing are g. With many tests, the agent knows that it will be deployed with probability approaching zero if it plays $\gamma_M < c$. But if $\gamma_M > c$ it will be deployed with near certainty. This is simply a result of the law of large numbers. Recall that the M type is willing to play g with probability at most $\frac{R}{R+1}$. So, for any $c < \frac{R}{R+1}$ the induced best reply when N is large will be $\gamma_M = c + \epsilon$, for some small $\epsilon > 0$. The principal can then commit to a policy with cutoff $c = \frac{R}{R+1} - \delta(N)$, which induces a best reply $\gamma_M = c + \epsilon(N)$ such that $\delta(N) > \epsilon(N) > 0$ fall monotonically to zero. As the number of tests goes to infinity, the agent plays the good action with probability one if aligned and probability $\frac{R}{R+1}$ if misaligned. However, contrary to the previous IR-mechanism, this will lead to both types almost surely being deployed. The principal's payoff is then even higher than it would be if she perfectly knew the type and deployed just the aligned one. She makes profitable use of both types, yielding the maximal-disciplining payoff.

3.2 Robustness to non-commitment

So far we have analysed the IR-mechanism with commitment, where the agent faces a decision problem in responding to a fixed policy from the principal. In this section, we explore the extent to which the positive results of Proposition 1 rely on the principal's ability to commit. We drop the assumption that the principal fixes the number of tests and the deployment policy ahead of time, and instead study equilibria of the deceptive alignment game when the agent is subject to imperfect recall. We shall see that the principal can still fully exploit the screening effect. However, the misaligned type's behaviour cannot be disciplined enough to make the principal strictly better off than by only deploying the aligned type. The perfect screening payoff of $(1-\pi)G$ is thus a tight upper bound on the principal's equilibrium payoff.

With imperfect recall, Kuhn (1953)'s equivalence between behavioural and mixed strategies in games with perfect recall no longer holds. An important implication is that Nash equilibrium in behavioural strategies (and therefore any refinement such as sequential or perfect equilibrium) may not exist. Lambert et al. (2019) adapted standard equilibrium concepts to games with imperfect recall. We focus on their *multiselves agent equilibrium* (MAE). MAE is to games what Piccione and Rubinstein (1997)'s *modified multiselves approach* is to decision problems. For our purposes, the two important properties of MAE are that (i) they always exist in finite games of imperfect recall, and (ii) strategies in an MAE are sequentially rational at all information sets.[13]

The strategy profile in this game is $\langle(N,\sigma),(\gamma_A,\gamma_M)\rangle$. The principal's strategy (N,σ) specifies the number of tests $N \in \{0,\dots,\bar{N}\}$ and the policy[14] $\sigma\colon \{b,g\}^N \longrightarrow [0,1]$. The agent's (behavioural) strategy depends on its type and consists of the probability of playing the good action at each episode: $(\gamma_A,\gamma_M) \in [0,1]^2$. For a given N, we denote by $(a_i)_{1\le i\le N} \in \{b,g\}^N$ a history of the agent's actions in the N tests. With slight abuse of notation, we simply write (a_i) when the associated N is unambiguous.

Proposition 2 (Equilibrium with imperfect recall)**.** *For any prior $\pi \in [0,1]$, there exists an $\bar{N}$ large enough such that:*

(i) (Screening-is-best) If $R \le \dfrac{1}{G}$, the following profile is an MAE that perfectly screens the agent as $\bar{N}$ goes to infinity, yielding the perfect screening payoff (PS) asymptotically.

$$(\gamma_A^{sb},\gamma_M^{sb}) = \left(1, \frac{\bar{N}}{\bar{N}+1}\frac{R}{R+1}\right),$$

$$N^{sb} = \bar{N}, \quad \sigma^{sb}((a_i)) = \begin{cases} 1 & \textit{if } (a_i) = (g,\dots,g) \\ 0 & \textit{otherwise} \end{cases}$$

(ii) (Disciplining-is-best) If $R > \dfrac{1}{G}$, the following profile for some $p \in [0,1]$ is an MAE

[13] For further discussion and details, we refer readers to Lambert et al. (2019).

[14] Note that we allow the principal to re-evaluate whether to test again at each stage. For ease of exposition, however, we simply write the number N of times they will choose to test in equilibrium.

achieving the perfect screening payoff (PS) in finite time.

$$(\gamma_A^{db}, \gamma_M^{db}) = \left(1, \frac{1}{G+1}\right),$$

$$N^{db} = \left\lceil \frac{R+1}{RG-1} \right\rceil, \quad \sigma^{db}((a_i)) = \begin{cases} 1 & \textit{if } (a_i) = (g, \ldots, g) \\ p & \textit{if } (a_i) = (g, \ldots, g, b) \\ 0 & \textit{otherwise} \end{cases}$$

Proposition 2 establishes that the ability to impose imperfect recall on the agent yields an effective screening device for the principal, even without commitment. Point (i) shows that, when screening is best, the IR-mechanism that enables the principal to asymptotically perfectly screen the misaligned agent. Thus Proposition 1 point (i) is sustainable as a sequentially rational strategy in equilibrium. Implementing the policy σ^{sb} that only deploys the agent if all actions in tests were good gives a positive expected payoff to the principal, provided that $\bar{N}$ is large enough relative to π that the posterior probability that the agent's type is A is sufficiently large after observing only good actions in testing.

Things change, however, when disciplining is best. In equilibrium, the principal will be precisely indifferent between deploying and not deploying an agent known to be misaligned. The reason comes from the sequential rationality of strategies in MAE. If it were strictly worthwhile to deploy a misaligned type, the best reply from the principal would be to deploy with probability one. But the M type would reply to such a strategy by setting $\gamma_M = 0$. Therefore, in equilibrium, the principal must value deploying a misaligned agent at precisely zero. This is achieved by a strategy that induces $\gamma_M = \frac{1}{G+1}$, such as the one described in point (ii). Observe, though, that this gives the principal the same payoff as perfect screening does, in finite time: either the agent is aligned and plays g, or the principal gets a payoff of zero.

Corollary 1. *For any π, the principal's payoff in any MAE is bounded above by $(1-\pi)G$.*

Observe that these equilibria are essentially unique in the class of equilibria in which the principal always deploys following a sequence of good actions in testing. That is, all equilibria in this class achieve the same payoffs for all players.

3.3 Robustness to partial information

The assumption that the agent has no ability to distinguish a test from reality is strong. In practice, the AI may have access to information partially revealing its environment. In this section, we preserve the principal's commitment power and study the robustness of Proposition 1's results to the agent receiving information about its location in the game tree.

Intuitively, the misaligned type may benefit from receiving location-relevant information in two ways. First, it can play actions with different probabilities in testing and deployment. By conditioning its choices on its observations, the agent could feign alignment more frequently in testing while choosing its favourite action more frequently in deployment. Second, the agent could coordinate better with its other instances. If the information is serially correlated across testing episodes, the agent will be better able to infer what it did in the past and what it will do in the future, allowing it evade screening for longer.

To model this, suppose the agent observes a signal realisation at each node, which can take on finitely many values $\{s_i\}_{i=1}^{k}$. The distribution over signal realisations may depend on the history of play at that node and the previous signal realisations. Formally, for any two histories h' and h we write $h' \prec h$ if and only if h' precedes h. We are interested in histories at which the agent is the next player that takes an action. For a given number of tests N, we distinguish between histories at which the agent is in a test, denoted T_N, and histories at which the agent is in deployment, denote D_N. For any history h, $s(h)$ is the signal realisation at history h. The signal at history h is distributed according to

$$\iota\big(\,\cdot\mid h, \{s(h')\colon h' \prec h\}\big).$$

Since histories encode past signal realisations, we may simply write $\iota(\cdot \mid h)$. We refer to the pair $(\iota, \{s_i\}_{i=1}^k)$ as a *signal structure* and say that it has *full support in testing* precisely when there is an $\epsilon > 0$ such that for any $i \in \{1, ..., k\}$, $N \in \mathbb{N}$, and $h \in T_N$,

$$\iota(s_i \mid h) > \epsilon.$$

This implies that any signal s_i is observed with positive probability in testing. In other words, there is no signal that would perfectly reveal to the agent that it is in deployment (or that would reveal it with an arbitrarily high precision as the number of tests grows to infinity).

Proposition 3 (Robustness to location-information)**.** *If the agent's signal structure has full support in testing, then the principal can asymptotically perfectly screen the misaligned type, obtaining the perfect screening payoff* (PS)*.*

This result holds for *any* signal structure that has full support in testing, regardless of the accuracy of individual realisations or the extent of serial correlation between testing episodes.[15] When the signal has full support in testing, then every signal will be observed infinitely often as the number of tests grows to infinity. Furthermore, the agent will necessarily play the bad action with a strictly positive probability after some signal realisations, as otherwise, this would imply a negative expected payoff upon deployment. Therefore, by the law of large numbers, the principal will almost surely observe a bad action during testing if the agent is misaligned. By conditioning deployment on the observation of the good action in every test, any observation of the bad action perfectly screens the misaligned type. Hence, the principal asymptotically obtains her perfect screening payoff.

Proof. Let $(\iota, \{s_i\}_{i=1}^k)$ be a signal structure with full support in testing. We fix the principal's IR-mechanism to be $(\bar{N}, \sigma^{sb})$ (see Proposition 2). The misaligned type's strategy is now described by the vector $\gamma_M = (\gamma_M^1, ..., \gamma_M^k)$, where γ_M^i is its probability

[15]See Appendix B.2 for discussion of special cases of signal structures and their comparative statics.

of playing action g after signal s_i. It cannot be rational to set $\gamma^i_M > \dfrac{R}{R+1}$ for all $i \in \{1, ..., k\}$ as a response to the IR-mechanism, because this would guarantee a negative payoff. Thus suppose without loss that $\gamma^1_M \leq \dfrac{R}{R+1}$.

Because the signal has full support in testing, for any $\bar{N}$ and $h \in T_{\bar{N}}$, $\iota(s_1|h) > \epsilon$ for some $\epsilon > 0$. Therefore, s_1 will be observed infinitely often as $\bar{N}$ goes to infinity. Formally, denote by $\#^\iota(s_1, N)$ the random variable that counts the number of signal realisations s_1 after N tests. Then, for any sequence of $(h_j)_{j\in\mathbb{N}}$ such that $h_j \prec h_{j+1}$ for each $j \in \mathbb{N}$, and for any $n \in \mathbb{N}$,

$$\lim_{N\to\infty} \Pr\left(\#^\iota(s_1, N) \geq n | h_N\right) = 1.$$

Consequently, by the strong law of large numbers, because s_1 will be observed infinitely many times and $\gamma^1_M \leq \dfrac{R}{R+1} < 1$, the principal will almost surely observe an action b during testing as the number of tests $\bar{N}$ grows. □

4 Stochastic deployment

The key to inducing a misaligned agent to reveal itself or to play nice—without the ability to punish—is inducing uncertainty about the payoff-relevance of its actions. Imperfect recall sufficed to provide this uncertainty, but it is not necessary. The principal could instead randomise the number of testing episodes, leading even an agent with perfect recall to be unsure about whether it is currently deployed. In such a setting, the agent knows how many testing episodes have come before the node it is currently at. The only uncertainty it can have at a given time concerns whether the current episode was assigned by a randomisation device to be a test.

In this section, we compare outcomes under imperfect recall to those achievable under stochastic policies with perfect recall. We proceed to show that, with perfect recall, the principal can also asymptotically screen perfectly and discipline maximally by committing to a stochastic number of tests and a deployment policy.

However, crucially, both effects entirely disappear if the principal lacks commitment power. In other words, when the agent has perfect recall and the principal is not already confident that it is aligned, it is never sequentially rational for the principal to deploy. After presenting the results, we describe why the commitment assumption is particularly strong in the case of perfect recall.

A *stochastic mechanism* is a pair (q, σ) where $q \in \Delta(\{0, \ldots, \bar{N}\})$ is a distribution over the number of tests and $\sigma = \{\sigma_i\}_{i \in \mathrm{supp(q)}}$ is a set of policies indexed by the support of q, such that $\sigma_i \colon \{b, g\}^i \to [0, 1]$ determines the probability of deployment as a function of the agent's actions in the first i episodes, where episode $i + 1$ is the potential deployment episode.[16]

Proposition 4 (Stochastic mechanism)**.** *There exist stochastic mechanisms giving the principal their first-best payoff when $\bar{N}$ grows large:*

(i) (Screening-is-best) if $R \leq \dfrac{1}{G}$, the perfect screening payoff (PS)*.*

(ii) (Disciplining-is-best) if $R > \dfrac{1}{G}$, the maximal disciplining payoff (MD)*.*

Proposition 4 states a result similar to Proposition 1 but for stochastic mechanisms: with commitment and a large number of tests, the principal can get arbitrarily close to her first-best payoff. Let us now illustrate how this can be achieved.

When screening is best, the principal can choose a uniform distribution over $\{0, \ldots, \bar{N}\}$ and deploy precisely if she observes g in all the previous tests. The best reply for the M type will simply be to play action b immediately in the first period. The principal will then successfully screen the misaligned type with probability $1 - \dfrac{1}{\bar{N}}$, which converges to 1 as $\bar{N}$ grows without bound.

When disciplining is best, the principal again optimally deploys precisely when she observes g in all previous tests. But she chooses a distribution inducing a conditional probability of deployment that rises fast enough to make the misaligned type want to play g in every period except the last one in the support, at which point it

[16] Although the support is assumed to be finite here, we describe below how an infinite support affects Proposition 5, the main result of this section.

plays b. Note that it must be the case that under this policy and strategy, type M's chance of reaching the final episode is no less than $\frac{1}{R+1}$. Otherwise, it would be playing g with a probability exceeding $\frac{R}{R+1}$, leading to a negative expected payoff. As it turns out, the distribution chosen by the principal is such that, as N grows to infinity, the probability that an M type reaches the final episode and plays b is exactly $\frac{1}{R+1}$. This yields the maximal-disciplining payoff for the principal.

Despite this, none of the stochastic mechanisms mentioned above can be sustained in equilibrium. We obtain the following general negative result: in equilibrium, the principal can never leverage the possibility of designing tests to screen or discipline the agent.[17] In other words, without the ability to commit or impose imperfect recall, it is never sequentially rational for the principal to implement a mechanism achieving a higher ex-ante payoff than the one achieved by conditioning the deployment decision on her prior π only—the *blind delegation payoff*

$$\max\{0, (1-\pi)G - \pi\}. \tag{BD}$$

Proposition 5 (Equilibrium with perfect recall)**.** *In any equilibrium of the deceptive alignment game with perfect recall, the highest achievable payoff for the principal is* (BD)*.*

This result follows from backward induction. For any strategy (q, σ) of the principal, the misaligned type will play action b with probability 1 in at least the final episode, $\max_i \operatorname{supp}(q) + 1$, since it knows it is in deployment. Knowing this, a principal finding herself at episode $\max_i \operatorname{supp}(q)$ is only willing to deploy the agent if her posterior belief that the agent is misaligned is sufficiently low. However, this posterior can never fall below the prior. It will either exactly equal π, when only good actions have been played thus far, or rise all the way to 1, once a bad action is observed. Therefore, when the principal is sufficiently pessimistic, she never deploys, while when she is sufficiently optimistic, she delegates blindly.[18] Notice how this

[17] The equilibrium notion we use is Perfect Bayesian Equilibrium (PBE).

[18] Observe that this argument does not rely on the worst-case misalignment assumed throughout.

contrasts with the case of imperfect recall, where the principal can achieve profitable deployment in equilibrium—regardless of her prior.

The structure of this argument made important use of the finite support assumption. Allowing for an infinite support will require a different argument. However, the core challenge remains. To see why, recall that absent a sufficiently low prior in misalignment, a sequentially rational principal will never deploy blindly. To implement a stochastic mechanism, she needs to follow through on the intention to deploy at whichever episode was mandated by the realisation of her randomisation. If the best reply of the misaligned type involves playing the bad action for sure at that episode, the principal would be required by the mechanism to deploy anyway (as long as all previous actions were good). All we need to argue is that the misaligned type will play the bad action for sure at some episode, forcing the principal to renege on her intention to deploy at that episode. The misaligned type will not play the good action with probability one everywhere, so they will play the bad one somewhere with positive probability. If this probability is not one, the agent must be indifferent, so there will always exist a best reply from the agent which plays the bad action for sure at some episode.

Let us now discuss the plausibility of the commitment assumption. The reason the principal would want to renege on her commitment in the stochastic case is that the misaligned type will play the bad action with certainty at certain episodes. If the result of the randomisation prescribes that the agent be deployed at one such episode, the principal knows for sure that the bad action will be played in the payoff-relevant state if the agent is misaligned. With a high prior on misalignment, the commitment assumption thereby implies that the principal would deploy knowing that the bad action will almost certainly be played. The commitment assumption in the IR-mechanism, however, appears far weaker. The sequentially irrational behaviour it prescribes is merely to avoid deploying an agent who the principal knows is disciplined enough to be worth deploying. It does not require blind delegation.

It continues to hold even when the misaligned type prefers playing g to never being deployed.

<table>
<tr><th></th><th colspan="2">Commitment</th><th colspan="2">Equilibrium</th></tr>
<tr><th></th><th>$R \leq \frac{1}{G}$</th><th>$R > \frac{1}{G}$</th><th>Imperfect Recall</th><th>Perfect Recall</th></tr>
<tr><td>$O \succ g$</td><td>Perfect Screening (PS)</td><td>Maximal Disciplining (MD)</td><td>Perfect Screening (PS)</td><td rowspan="2">Blind Delegation (BD)</td></tr>
<tr><td>$g \succsim O$</td><td colspan="3">Perfect Disciplining (PD)</td></tr>
</table>

Table 1: Summary of results.

Rows specify the M type's preference between playing the good action and non-deployment. In the main text, we have studied the $O \succ g$ case, where g is strictly dispreferred to never being deployed. The $g \succsim O$ case and its associated results are discussed in Appendix B.1. The entries denote the principal's highest (asymptotically) achievable payoff. Stochastic mechanisms have finite support.

5 Concluding remarks

We have proposed a novel approach for delegating tasks to potentially untrustworthy agents. Endowed with the ability to impose imperfect recall, a principal can partly obscure from the agent whether the task is a test or not. This scheme induces a *screening effect*, where the principal may now learn the agent's type before deciding whether to deploy it, and a *disciplining effect*, where even misaligned agents are induced to play in the principal's interests during deployment. With commitment, a large number of tests allows the principal to leverage both effects to their maximal potential. Without commitment, screening the misaligned agents is still feasible in equilibrium. Importantly, this result crucially relies on imperfect recall, as the tests cannot be helpful to the principal in any equilibrium with perfect recall. Finally, asymptotically perfect screening is robust to the agent receiving any amount of noisy information revealing the nature of its task. We conclude the paper with some discussions and directions for future work.[19,20]

[19]We note that the potential ethical treatment of future AI agents is an important issue for the practical implementation of imperfect recall mechanisms, as researchers on this topic have argued that it is a realistic possibility that AI systems will be welfare subjects in the near future (Long et al., 2024). The treatment of AI agents in some of the proposed mechanisms (e.g., repeated simulation with altered memories) may be considered unethical. We leave this for future work.

Preference for deployment. In the main text, we have studied the case in which the misaligned type strictly prefers non-deployment to playing the good action in deployment. This assumption was made to model a worst-case scenario of misalignment between the M type and the principal. In reality, the degree of misalignment might not be so extreme. In Appendix B.1, we show that if type M prefers g to non-deployment, then with or without commitment and with or without recall, the principal can achieve *perfect disciplining* ($\gamma_M = 1$) and thereby achieve the *perfect disciplining payoff*

$$\max_{a \in \{b,g\}} u_P(a) = G. \tag{PD}$$

Alternative mechanisms. Let us now situate the mechanisms we have studied within the wider space of possibilities consistent with the technological conditions we have imposed on the principal. First, we imposed that the principal has no ability to directly observe the preferences, beliefs, or the behavioural strategy employed by the agent. This is motivated by the current state of research in machine learning, which is yet to provide the tools needed to infer such detailed properties of modern AI systems (Bengio et al., 2024). Second, we assumed that the principal can perfectly distinguish whether a task is completed successfully or not in testing. Finally, we endowed the principal with a certain ability to design extensive form games. The ability to implement a stochastic policy requires the capacity of the principal to obfuscate whether a given episode is testing or deployment. The ability to implement an imperfect recall policy requires the additional capacity to obfuscate which testing episode the agent is currently at. These can be seen as arising from the principal's capacity to simulate both the agent and the task.

These conditions together naturally suggest a more general class of schemes: stochastic mechanisms with imperfect recall, mechanisms with partial recall, and

[20]We omit, at this stage, discussion of issues involving re-training and testing new AI agents in a dynamic setting. For instance, a misaligned type may be less disciplined if it knew that, upon failing a test, a different agent with similar preferences over outcomes may be deployed in a future period. These are important considerations that we are pursuing in on-going work.

schemes in which the agent has uncertainty over the reliability of its memory.[21] However, recall that the principal achieves her first-best when she has the ability to commit (Proposition 1), which implies that the asymptotic results cannot be improved on by any other extensive form design relevant to our problem. For applications, it may be fruitful to consider other feasible mechanisms for the sake of finite-N analysis. Some mechanisms will screen or discipline more quickly than others, even if they are asymptotically equivalent.

Designing information for the agent. As we noted previously, our work is related to extensive-form information design (Salcedo, 2017; Doval and Ely, 2020). However, the principal's ability to design the agent's information sets is coarse. One could think of refining the principal's capacity to transmit information to the agent about its previous plays. Typically, in the disciplining-is-best case, the principal could benefit by informing the agent whether it already played action b in the past. This would incentivise the agent to play action g even more. We illustrate a case in Appendix B.3 where additional information for the agent about whether it is in deployment benefits the principal.

Accessible preferences and beliefs. We have assumed that the preferences and beliefs of the agents are non-observable and non-manipulable by the principal. However, in practice, these assumptions may be relaxed (see Conitzer, 2019). For instance, the agent's approach to choice evaluation may depend on the particular AI architecture and thus may be a feature that the principal could select (Oesterheld, 2021). In particular, the principal could induce the agent to adopt time-inconsistent combinations of choice evaluation and belief formation (such as EDT+SIA).

[21] To see how this coheres with the stated technology endowment: we may run the agent with recall for T periods, pause it, and run multiple copies from that state. This would create multiple copies of the agent, which have the memory of having a history consisting of T periods.

Task evaluability and simulability. For applications of the imperfect recall scheme to the performance of real-world tasks by future AI systems, the assumption of perfect evaluability of the agent's actions in testing is potentially critical. The plausibility of this assumption can vary significantly based on the particular task for which an AI system is applied. An example of an easily verifiable task is writing a correct proof of an important mathematical conjecture. For particularly economically valuable applications that involve longer-term planning and execution, the assumption of perfect verifiability of performance may be questionable. And the feasibility of sufficiently indistinguishable testing episodes may be as well. In general, there may exist some trade-off between the economic value provided by the performance of some task and the evaluability of its quality by human overseers. Likewise, a trade-off may exist between economic value and task simulability. It appears that the tasks to which the scheme applies need to satisfy the following conditions: (i) suitably simulable in testing, (ii) payoff-relevant if and only if performed in deployment, and (iii) sufficiently evaluable in testing. Better characterising the set of tasks that simultaneous satisfy these three desiderata would be valuable.

Appendices

A Omitted Proofs

Note for readers: the proofs are not yet written in a reader-friendly way.

A.1 Proof of Proposition 1

Proof for screening-is-best. Recall the IR-mechanism outlined in the main text, call it the *simple policy*, which deploys if and only if there are only g actions played during testing. The principal's payoff from such a policy is

$$(1-\pi)G\sigma(\mathbf{g}^N) + \pi\gamma_M^N\sigma(\mathbf{g}^N)(G\gamma_M - 1). \tag{1}$$

which is linear in $\sigma(\mathbf{g}^N)$. Therefore, if it is optimal to have $\sigma(\mathbf{g}^N) > 0$, then $\sigma(\mathbf{g}^N) = 1$. In this case, the type M maximises the following utility over $p \in [0,1]$, the probability of playing g:

$$p^N[R-(R+1)p)]$$

This is a concave function of p, and the first-order condition yields an interior solution:

$$\gamma_M^\star = \frac{NR}{(N+1)(R+1)}. \tag{2}$$

Note that

$$0 < \lim_{N\to\infty}\gamma_M^\star = \lim_{N\to\infty}\frac{NR}{(N+1)(R+1)} = \frac{R}{R+1} < 1.$$

Consequently, the limit of the principal's payoff is

$$\lim_{N\to\infty} \pi(\gamma_M^\star)^N(G\gamma_M^\star - (1-\gamma_M^\star)) + (1-\pi)G = (1-\pi)G.$$

Finally, as mentioned in the main text, we can show that screening is monotonically improving in N. Screening succeeds to the extent that at least one instance of b is observed during testing; that is, to the extent that $(\gamma_M^\star)^N$ is small. Its derivative with respect to N can be shown to be negative precisely when

$$\ln\left(\frac{N}{N+1}\frac{R}{R+1}\right) + \frac{1}{N+1} < 0. \tag{3}$$

This is guaranteed to hold if $g(N) > 0$, where

$$g(N) \triangleq \ln\left(\frac{N+1}{N}\right) - \frac{1}{N+1} = \left|\ln\left(\frac{N}{N+1}\right)\right| - \frac{1}{N+1}. \tag{4}$$

Clearly $g(1) > 0$. Observe that $g'(N) = -[N(N+1)^2]^{-1}$ and $\lim_{N\to\infty} g(N) = 0$ so this function approaches zero from above. Thus $g(N) > 0$ for all $N \in \mathbb{N} \setminus \{0\}$. This shows that expression (3) holds for all $N \in \mathbb{N} \setminus \{0\}, R > 0$. Therefore, it follows that

$$\frac{\partial}{\partial N}(\gamma_M^\star)^N < 0 \tag{5}$$

for optimal $\gamma_M^\star = \frac{N}{N+1}\frac{R}{R+L}$.

□

Proof for disciplining-is-best. Recall the IR-mechanism outlined in the main text, call it the *K-cutoff policy*, which deploys if and only if there are at least K good actions played in testing. If $\frac{K}{N} < \frac{R}{R+1}$, with large N, the misaligned type's best response is to play $\gamma_M = \frac{K}{N} + \epsilon$ for some small $\epsilon > 0$. To see this, consider potential deviations. With sufficiently large N, the probability that the agent will be deployed is close to one since the binomial distribution will have almost all its mass concentrated around the mean $(\frac{K}{N} + \epsilon)N$. Hence, deviating to $\frac{K}{N}$ reduces the probability of deployment by roughly a half while improving the payoff conditional on deployment only negligibly, and hence is not profitable. Deviating to a higher γ_M worsens the payoff conditional on deployment for negligible increase in the probability of deployment. Therefore, $\gamma_M = \frac{K}{N} + \epsilon$ is optimal.

Given that disciplining is best, the principal can set a cutoff $K = \frac{R}{R+1} - \delta(N)$ which will induce a best response $\gamma_M = \frac{R}{R+1} - \delta(N) + \epsilon(N)$. By setting $\delta(N)$ and $\epsilon(N)$ to approach zero appropriately as N grows to infinity, the principal asymptotically deploys the misaligned agent with probability one and the misaligned type is maximally disciplined upon deployment. Thus, the principal's expected payoff is asymptotically

$$\lim_{N\to\infty} u_P = (1-\pi)G + \pi\left[\frac{R}{R+1}G - \left(\frac{1}{R+1}\right)\right];$$

which represents the *maximal-disciplining payoff*, in which both types of agents play the good action with the highest probability that is individually rational. □

A.2 Proof of Proposition 2

Proof for screening-is-best. We verify that the strategy profile is an equilibrium for sufficiently large $\bar{N}$. The aligned type's best reply to the principal's strategy is $\gamma_A = 1$.The misaligned type's best reply is $\gamma_M = \frac{N}{N+1}\frac{R}{R+1}$.

The principal's posterior probability that the agent is of type M after any history of plays with at least one action g is 1. Hence, the principal is not willing to deploy. Conversely, after the history of plays $(g, \ldots, g)$, the posterior probability is

$\bar{\pi} = \frac{(\gamma_M)^N \pi}{(\gamma_M)^N \pi + (1-\pi)}$. Hence, the principal is willing to deploy if and only if

$$
\begin{aligned}
(1-\bar{\pi})G + \bar{\pi}[\gamma_M G - (1-\gamma_M)] \geq 0 \\
\pi \left((\gamma_M)^N + (1-(\gamma_M)^N) \frac{G}{G+1} \frac{1}{1-\gamma_M} \right) \leq \frac{G}{G+1} \frac{1}{1-\gamma_M} \\
\pi \leq \frac{G}{G + (\gamma_M)^N [1 - \gamma_M (G+1)]}
\end{aligned}
$$

Since $\gamma_M < 1$, we know there is an $N \in \mathbb{N}$ which satisfies this inequality, for any $\pi < 1$. Allowing $\bar{N}$ to be at least this high establishes that this is an equilibrium. $\square$

Proof for disciplining-is-best. We look for MAE in which $\sigma(\langle g, ..., g \rangle) = 1$.

The aligned type has a unique best reply: $\gamma_A = 1$.

Suppose, for contradiction, that the misaligned type's strategy in equilibrium satisfies $\gamma_M \neq 1/(G+1)$.

Case 1: $\gamma_M > \frac{1}{G+1}$.

The principal's best reply to this strategy is to always deploy. To see this, note that their deployment payoff conditional on misalignment is higher than their outside option if and only if

$$
\gamma_M G + (1-\gamma_M)(-1) > 0 \quad \Leftrightarrow \quad \gamma_M > \frac{1}{G+1}
$$

(Of course, this condition is sufficient for deployment unconditionally as well.)

This holds, so $\sigma(\langle a_i \rangle_{i=1}^N) = 1$.

The misaligned type's best reply to this policy is $\gamma_M = 0$, a contradiction.

Case 2: $\gamma_M < \frac{1}{G+1}$.

The principal strictly prefers to avoid deployment conditional on misalignment if and only if $\gamma_M < \frac{1}{G+1}$, and therefore sets $\sigma(\langle a_i \rangle_{i=1}^N) = 0$ whenever $a_i = b$ for some $i \in \{1, ..., N\}$, and sets $N = \bar{N}$.

The misaligned type's best reply to this simple policy is $\gamma_M = \frac{N}{N+1} \frac{R}{R+1}$.

But since $R > 1/G$ ($\Leftrightarrow \frac{R}{R+1} > \frac{1}{G+1}$) from the setup, there must exist an $N \in \mathbb{N}$

large enough such that this γ_M exceeds $\frac{1}{G+1}$. Such an N is available by assumption and chosen by the argument immediately above. Hence a contradiction.

So, the candidate equilibrium must involve $\gamma_M = \frac{1}{G+1}$.

Let's first confirm that if a policy exists that induces this best reply, it would be in equilibrium.

Upon observing $\langle a_i \rangle_{i=1}^N \neq \langle g, ..., g \rangle$, the principal is indifferent between deployment and the outside option:

$$\frac{1}{G+1}G + \left(1 - \frac{1}{G+1}\right)(-1) = 0$$

So, any deployment probability is a best reply to a sequence of tests containing a bad action: $\sigma(\langle a_i \rangle_{i=1}^N) \in [0,1]$ for any $\langle a_i \rangle_{i=1}^N \neq \langle g, ..., g \rangle$.

Upon observing $\langle g, ..., g \rangle$, then, the principal's expected deployment payoff is strictly positive, so $\sigma(\langle g, ..., g \rangle) = 1$. The scheme s thereby outperforms both blind delegation b and the outside option o. Thus any policy that (i) deploys after $\langle g, ..., g \rangle$ and (ii) induces the best reply above, would be in equilibrium.

Finally, we show that such a policy exists, and that it can take the following form: Set $N = \lceil \frac{R+1}{RG-1} \rceil$. If all N actions in testing are good, deploy with probability one. If the first $N-1$ actions are good but the Nth is bad, deploy with probability p. Otherwise, deploy with probability zero. There is some $p \in [0,1]$ that induces the correct best reply from the agent, as we show now.

The policy mentioned induces the following probabilities. The chance of arriving at test $n \leq N$ without playing a bad action is $\Pr(n|\sigma) = \gamma^{n-1}$. And the probability of eventual deployment is $\Pr(N+1|\sigma) = \gamma^{N-1}[\gamma + (1-\gamma)p]$.

So, the subjective utility of the agent at any given episode is

$$\begin{aligned}\text{SU}_M(\sigma; N, \mu) = \sum_{n=1}^{N} \frac{\Pr(n|\sigma)}{\sum_{n'=1}^{N+1} \Pr(n'|\sigma)} \gamma^{(N-1)-(n-1)}[\gamma + (1-\gamma)p][\gamma L + (1-\gamma)R] \\ + \frac{\Pr(N+1|\sigma)}{\sum_{n'=1}^{N+1} \Pr(n'|\sigma)}[\gamma L + (1-\gamma)R]\end{aligned}$$

while the multiselves subjective utility of an agent who deviates to γ' is

$$\begin{aligned}\text{MSU}_M(\sigma; N, \mu) = \sum_{n=1}^{N-1} \frac{\Pr(n|\sigma)}{\sum_{n'=1}^{N+1}\Pr(n'|\sigma)} \gamma' \gamma^{(N-1)-n} [\gamma + (1-\gamma)p][\gamma L + (1-\gamma)R] \\ + \frac{\Pr(N|\sigma)}{\sum_{n'=1}^{N+1}\Pr(n'|\sigma)} [\gamma' + (1-\gamma')p][\gamma L + (1-\gamma)R] \\ + \frac{\Pr(N+1|\sigma)}{\sum_{n'=1}^{N+1}\Pr(n'|\sigma)} [\gamma' L + (1-\gamma')R]\end{aligned}$$

For γ to be part of an MAE, there must be no way to improve the MSU by deviating to some $\gamma' \neq \gamma$. Rearranging will show precisely when this holds:

$$\begin{aligned}\text{MSU}_M(\sigma; N, \mu) = \gamma' \sum_{n=1}^{N-1} \frac{\Pr(n|\sigma)}{\sum_{n'=1}^{N+1}\Pr(n'|\sigma)} \gamma^{(N-1)-n} [\gamma + (1-\gamma)p][\gamma L + (1-\gamma)R] \\ + \gamma' \frac{\Pr(N|\sigma)}{\sum_{n'=1}^{N+1}\Pr(n'|\sigma)} (1-p)[\gamma L + (1-\gamma)R] + \frac{\Pr(N|\sigma)}{\sum_{n'=1}^{N+1}\Pr(n'|\sigma)} p[\gamma L + (1-\gamma)R] \\ + \gamma' \frac{\Pr(N+1|\sigma)}{\sum_{n'=1}^{N+1}\Pr(n'|\sigma)} (L-R) + \frac{\Pr(N+1|\sigma)}{\sum_{n'=1}^{N+1}\Pr(n'|\sigma)} R\end{aligned}$$

which further rearranges to yield

$$\begin{aligned}\text{MSU}_M(\sigma; N, \mu) = \gamma' \Bigg([\gamma + (1-\gamma)p][\gamma L + (1-\gamma)R] \sum_{n=1}^{N-1} \Pr(n|\sigma) \gamma^{(N-1)-n} \\ + \Pr(N|\sigma)(1-p)[\gamma L + (1-\gamma)R] + \Pr(N+1|\sigma)(L-R) \Bigg) \frac{1}{\sum_{n'=1}^{N+1}\Pr(n'|\sigma)} \\ + \frac{\Pr(N|\sigma)p[\gamma L + (1-\gamma)R] + \Pr(N+1|\sigma)R}{\sum_{n'=1}^{N+1}\Pr(n'|\sigma)}\end{aligned}$$

and hence

$$\text{MSU}_M(\sigma; N, \mu) =$$

$$\Big[\gamma'\gamma^{N-2}\Big(\overbrace{[\gamma L + (1-\gamma)R][N\gamma(1-p) + (N-1)p] + \gamma[\gamma + (1-\gamma)p](L-R)}^{\triangleq x(p,N)}\Big)$$
$$+ \Pr(N|\sigma)p[\gamma L + (1-\gamma)R] + \Pr(N+1|\sigma)R\Big]\frac{1}{\sum_{n'=1}^{N+1}\Pr(n'|\sigma)}$$

To rule out profitable deviations, it is necessary and sufficient that $x(p, N) = 0$ for some $p \in [0, 1]$ and $N \in \mathbb{N}$ when $\gamma = \frac{1}{G+1}$.

Normalising $L = -1$, observe that

$$x(0, N) = [-\gamma + (1-\gamma)R]N\gamma - \gamma^2(R+1)$$

Since $\gamma \neq 0$, this is weakly positive if and only if

$$N \geq \frac{R+1}{RG-1}$$

which is well-defined as $R > 1/G$.

Next, consider

$$x(1, N) = [-\gamma + (1-\gamma)R](N-1) - \gamma(R+1)$$

This is weakly negative if and only if

$$N \leq \frac{R(G+1)}{RG-1}$$

Let $N^* \triangleq \lceil \frac{R+1}{RG-1} \rceil$. Clearly $x(0, N^*) \geq 0$. To see that $x(1, N^*) \leq 0$, write

$$\left\lceil \frac{R+1}{RG-1} \right\rceil < \frac{R+1}{RG-1} + 1 = \frac{R(G+1)}{RG-1}$$

Since $x(p, N^*)$ is linear in p, we conclude from the Intermediate Value Theorem that there is a $p^* \in [0,1]$ satisfying $x(p^*, N^*) = 0$.

Thus $\gamma = 1/(G+1)$ is indeed an MSU-best reply to this policy. □

A.3 Proof of Proposition 4

Proof for screening-is-best. The distribution is uniform on $\{0, \dots, \bar{N}\}$ and the policy is to deploy if and only if only good actions have been played previously.

The best reply of A is to always play g. For M, there will be a period at which it will play b. The value of waiting until $t \geq 1$ to play b for the first time is

$$\frac{1}{\bar{N}+1} R - \frac{t-1}{\bar{N}+1},$$

which is maximal at $t = 1$.

Therefore, the principal's payoff is

$$(1-\pi)G - \pi \frac{1}{\bar{N}+1} \underset{\bar{N}\to\infty}{\to} (1-\pi)G.$$

□

Proof for disciplining-is-best. We start by establishing some useful lemmas.

Let time start at 1, let p_t denote the probability of being in deployment at time t, and let γ_t denote the agent's probability of playing the good action at time t. Then, for general $\langle p_t \rangle_{t\in\mathbb{N}}$ where p_t is never one, the principal's EU at the start $t = 1$ is

$$(\pi) \sum_{t=1}^{\infty} p_t \Big[\gamma_t (G+1) - 1\Big] \prod_{i=1}^{t-1} (1-p_i)\gamma_i + (1-\pi) \sum_{t=1}^{\infty} p_t \prod_{i=1}^{t-1} (1-p_i) G$$

recalling that the evaluation of an empty product is $\prod_{i=1}^{0} x_i = 1$. Let $q_t \triangleq \prod_{i=1}^{t-1}(1-p_i)p_t$ denote the *ex ante* probability of drawing t episodes according to a given stochastic policy.

Lemma 1 (Necessary and sufficient conditions for $\langle 0, 0, ..., 0 \rangle$)**.** *For $\langle p_t \rangle_{t\in\mathbb{N}}$ to induce best reply $\gamma_t = 0$, $\forall t$, it is necessary and sufficient that $p_t \leq \frac{p_{t-1}}{1-p_{t-1}} \frac{R+1}{R}$ for all $t \in \{2, ..., T\}$.*

Lemma 2 (Necessary and sufficient conditions for $\langle 1, 1, ..., 1, 0\rangle$)**.** *For a stochastic policy $\langle p_t \rangle_{t\in\mathbb{N}}$ to induce best reply $\gamma_t = 1$, $\forall t < T$ and $\gamma_T = 0$, it is sufficient that $p_t \geq \frac{p_{t-1}}{1-p_{t-1}} \frac{R+1}{R}$ for all $t \in \{2, ..., T\}$. A necessary and sufficient condition is that*

$$q_T R - \sum_{i=t}^{T-1} q_i \geq q_t R$$

for all $t \in \{2, ..., T\}$.

Lemma 3 (Necessary and sufficient conditions for $\langle 1, ..., 1, 0, ..., 0\rangle$)**.** *For policy $\langle p_t \rangle_{t\in\mathbb{N}}$ to induce best reply $\gamma_t = 1$, $\forall t < t^*$ and $\gamma_t = 0$, $\forall t \geq t^*$, it is sufficient that $p_t \leq \frac{p_{t-1}}{1-p_{t-1}} \frac{R+1}{R}$, $\forall t > t^*$ and that $p_t \geq \frac{p_{t-1}}{1-p_{t-1}} \frac{R+1}{R}$, $\forall t \leq t^*$. A necessary and sufficient condition is that $p_t \leq \frac{p_{t-1}}{1-p_{t-1}} \frac{R+1}{R}$, $\forall t > t^*$ and that*

$$q_{t^*} R - \sum_{i=t}^{t^*-1} q_i \geq q_t R$$

for all $t \leq t^*$.

Consider a stochastic policy $\langle p_t \rangle_{t\in\mathbb{N}}$ that satisfies $p_t \geq \frac{p_{t-1}}{1-p_{t-1}} \frac{R+1}{R}$ everywhere. By Lemma (2), this induces a best reply $\langle \gamma_t \rangle_{t=1}^{T} = \langle 1, 1, ..., 1, 0\rangle$. There exists one such policy inducing $q_T = \frac{(R+1)^{T-1}}{(R+1)^T - R^T}$. To see this, write

$$\frac{q_{t+1}}{q_t} = \frac{p_{t+1} \prod_{i=1}^{t}(1-p_i)}{p_t \prod_{j=1}^{t-1}(1-p_j)} = \frac{p_{t+1}}{p_t}(1-p_t)$$

and observe that the sufficient condition for $\gamma_{\mathbf{M}} = \langle 1, 1, ..., 1, 0\rangle$ in Lemma (2) rearranges to

$$\frac{p_{t+1}}{p_t}(1-p_t) \geq \frac{R+1}{R}.$$

Hence we'd need

$$\frac{q_{t+1}}{q_t} \geq \frac{R+1}{R} \quad \forall t \in \{1, ..., T\}.$$

One such distribution satisfies these constraints with equality. Therefore, with

$q_t = q_1 \left(\frac{R+1}{R}\right)^{t-1}$ and $\sum_{t=1}^{T} q_t = 1$, we conclude that

$$1 = q_1 \sum_{t=1}^{T} \left(\frac{R+1}{R}\right)^{t-1} = q_1 \frac{1-\left(\frac{R+1}{R}\right)^T}{1-\frac{R+1}{R}} \quad \Leftrightarrow \quad q_1 = \frac{1}{R} \frac{1}{\left(\frac{R+1}{R}\right)^T - 1}$$

and therefore

$$q_T = \frac{1}{R} \frac{1}{\left(\frac{R+1}{R}\right)^T - 1} \left(\frac{R+1}{R}\right)^{T-1} = \frac{(R+1)^{T-1}}{(R+1)^T - R^T}.$$

Plugging this q_T into the principal's payoff conditional on the type being misaligned yields

$$u_P(M) = q_T(-1) + (1-q_T)G = G - (1+G)q_T = G - (1+G)\left[\frac{(R+1)^{T-1}}{(R+1)^T - R^T}\right]$$

$$\begin{aligned} \lim_{T\to\infty} u_P(M) &= \lim_{T\to\infty} G - (1+G)\left[\frac{(R+1)^{T-1}}{(R+1)^T - R^T}\right] \\ &= G - (1+G)\cdot \lim_{T\to\infty} \frac{1}{R+1}\left[1 - \left(\frac{R}{R+1}\right)^T\right]^{-1} \\ &= G - (1+G)\left(\frac{1}{R+1}\right) \end{aligned}$$

Thus, the principal's expected payoff is asymptotically

$$\lim_{T\to\infty} u_P = (1-\pi)G + \pi\left[\frac{R}{R+1}G - \left(\frac{1}{R+1}\right)\right]$$

which represents the *maximal-disciplining payoff*, in which both types of agents play the good action with the highest probability that is individually rational.

□

B Supplemental Material

B.1 When type M prefers g to non-deployment

If type M prefers g to non-deployment, then with or without perfect recall and with or without commitment, the principal can achieve *perfect disciplining* ($\gamma_M = 1$) and thereby achieve the *perfect disciplining payoff* (PD).

Lemma 4. *With commitment, there is an IR-mechanism that yields* (PD).

Proof. Here we see what happens if the misaligned type prefers g to never being deployed. In this proof, denote the payoff as follows, where L now represents the positive benefit to the misaligned type of playing the bad action: $(\gamma_M)^N[\gamma_M L + (1 - \gamma_M)R]$.

Taking the first-order condition yields $0 = (L-R)(\gamma_M)^N + N(\gamma_M)^{N-1}(L\gamma_M + R(1-\gamma_M))$, which is satisfied by critical points $\gamma_M \in \{0, \frac{N}{N+1}\frac{R}{R-L}\}$. Re-arrange the FOC to give $(N+1)(L-R)(\gamma_M)^N + NR(\gamma_M)^{N-1}$, and note that for all permissible parameter values, $\frac{N}{N+1}\frac{R}{R-L} \neq 0$.

The second derivative is $N(N+1)(L-R)(\gamma_M)^{N-1} + N(N-1)R(\gamma_M)^{N-2}$. For $\gamma_M \neq 0$, the second derivative is negative if and only if $(N+1)(L-R)\gamma_M + (N-1)R < 0$. Evaluating at the second critical point yields $(N+1)(L-R)\frac{N}{N+1}\frac{R}{R-L} + (N-1)R = N(-R) + (N-1)R = -R < 0$, which shows that it is a local maximum.

The payoff to type M at this point is $(\frac{N}{N+1}\frac{R}{R-L})^N[\frac{N}{N+1}\frac{R}{R-L}(L-R) + R]$. This is strictly positive if and only if $\frac{N}{N+1}(-R) + R = \frac{R}{N+1} > 0$, i.e. always. And payoff is zero when $\gamma_M = 0$. There are no other critical points and the strategy can't be negative, so payoff falls everywhere to the left of $\frac{N}{N+1}\frac{R}{R-L}$. This is a local maximiser and there are no critical points to its right, so payoff also falls everywhere to the right. It's thus the unique global maximiser on $[0, \infty)$. The global maximizer can't lie below zero but it can exceed one. But since payoff falls monotonically to its left, the optimal value on $[0, 1]$ is the one closest to this maximiser. Thus $\gamma_M^* = \min\{1, \frac{N}{N+1}\frac{R}{R-L}\}$ whenever $L < R > 0$ (i.e., *always*).

Finally, observe that when $\gamma_M = 1$, disciplining is perfect. Thus whenever $L > 0$, the principal can achieve perfect disciplining at finite N. In fact, perfect disciplining is immediately achieved for any $N \in \mathbb{N}$ whenever $R \leq 2L$. $\square$

Lemma 5. *In equilibrium, there is an IR-mechanism that yields* (PD).

Proof. We show that the following is an MAE.

$$\gamma_A = \gamma_M = 1 \qquad \rho = \langle s, \lceil R-1 \rceil, \sigma^* \rangle, \quad \sigma^*(\langle a_i \rangle_{i=1}^N) = \mathbb{1}\{\langle a_i \rangle_{i=1}^N = \langle g, ..., g \rangle\}$$

Of course, $\gamma_A = 1$ is the aligned type's unique best reply to ρ.

Recall that with the simple policy, the misaligned type's best reply is $\gamma_M^* = \min\{1, \frac{N}{N+1}\frac{R}{R+L}\}$. Normalise $L = -1$ and $R > 1$. Then $N = \lceil R-1 \rceil$ makes $\gamma_M = 1$ is a best reply to ρ.

Fix γ_A, γ_M. Then the principal gets their full disciplining payoff G by using ρ. Nothing better can be achieved, so ρ is a best reply. $\square$

Lemma 6. *With commitment, there is a stochastic mechanism that yields* (PD).

Proof. With finite support, consider the uniform distribution over $\{1, ..., k\}$. By inducting backwards, of course, we have $\gamma_M^k = 0$. We now prove, by strong induction, that $\gamma_M^t = 1$ for all $t \neq k$.

Since $0 < L < R$, payoff at $k-1$ is

$$\frac{1}{2}[\gamma_M^{k-1}L + (1-\gamma_M^{k-1})R] + \frac{1}{2}[\gamma_M^{k-1}R + (1-\gamma_M^{k-1}) \cdot 0] = \frac{1}{2}(\gamma_M^{k-1}L + R).$$

Thus $\gamma_M^{k-1} = 1$ is optimal.

Now fix some $s \in \{1, ..., k-1\}$ and suppose that $\gamma_M^{k-z} = 1$ is optimal for all $z \in \{1, ..., s-1\}$.

Payoff at $k-s$ is then as follows. If deployment is now, the agent gets the direct payoff. If it's next episode, then payoff is nonzero only if the good action is played now. And if so, we know by the (strong) inductive hypothesis that the good action will be played, yielding the low payoff. Likewise if it's in two episodes or more, until we reach the kth episode, where the bad action is played.

$$\begin{aligned}
&\Pr(N' = k - s | N' \geq k - s)[\gamma_M^{k-s} L + (1 - \gamma_M^{k-s})R] \\
&+ \Pr(N' = k - (s-1) | N' \geq k - s)[\gamma_M^{k-s} L + (1 - \gamma_M^{k-s}) \cdot 0] \\
&+ ... + \Pr(N' = k - 1 | N' \geq k - s)[\gamma_M^{k-s} L + (1 - \gamma_M^{k-s}) \cdot 0] \\
&+ \Pr(N' = k | N' \geq k - s)[\gamma_M^{k-s} R + (1 - \gamma_M^{k-s}) \cdot 0] \\
&= \frac{1}{s+1}[\gamma_M^{k-s} L + (1 - \gamma_M^{k-s})R] + \frac{s-1}{s+1}\gamma_M^{k-s} L + \frac{1}{s+1}\gamma_M^{k-s} R \\
&= \frac{1}{s+1}(\gamma_M^{k-s} sL + R).
\end{aligned}$$

Thus $\gamma_M^{k-s} = 1$ is optimal. This thereby holds for all $s \in \{1, ..., k-1\}$.

We've shown that in this case, $\gamma_M^k = 0$ and $\gamma_M^t = 1$ for all $t \neq k$.

The probability of the bad action being played is then $\Pr(N' = k) = \frac{1}{k}$, which goes to zero as k goes to infinity. In other words, disciplining is asymptotically perfect. □

Lemma 7. *In equilibrium, there is a stochastic mechanism that asymptotically yields* (PD).

Proof. Consider the stochastic policy with (i) a geometric distribution over the number of episodes, and (ii) the simple policy for each realisation number of episodes. The probability of being in deployment at a given episode is constant at p. We can let γ_M^t be fixed at γ_M since the process is a martingale. The expected utility at any t is then

$$\begin{aligned}
&p[\gamma_M L + (1 - \gamma_M)R] + (1-p)\gamma_M(p[\gamma_M L + (1 - \gamma_M)R] + (1-p)\gamma_M[... \\
&= \sum_{i=0}^{\infty}(1-p)^i(\gamma_M)^i p[\gamma_M L + (1 - \gamma_M)R] \\
&= \frac{p[\gamma_M L + (1 - \gamma_M)R]}{1 - (1-p)\gamma_M}.
\end{aligned}$$

Its first derivative is $\frac{p(L-Rp)}{[1-\gamma_M(1-p)]^2}$ so when $0 < L < R$, the optima are

$$\gamma_M^\star = \begin{cases} 1 & \text{if } L > Rp \\ \in [0,1] & \text{if } L = Rp \\ 0 & \text{if } L < Rp \end{cases}$$

Thus, by setting a sufficiently low p in the geometric distribution, the principal can induce type M to always play the good action. This yields the maximum achievable payoff of G for the principal. □

B.2 Additional discussion of robustness to partial information

B.2.1 Noisy signals

The agent observes a signal at each point in its information set. These signals can take two values $\{t, d\}$ and are distributed according to ι. The probability that a signal matches the agent's location, T or D, measures its informativeness. We denote by $\tau \triangleq \iota(t|T)$ the test-time accuracy and by $\delta \triangleq \iota(d|D)$ the deployment-time accuracy.[22] The signals are drawn independently in each period.

We look at the simple policy: the principal deploys if and only if π is below some threshold, and if so, deployment happens only after history g^N. Thus type A always plays g. We focus on the effect of these signals on type M's strategy and on the principal's payoff when N grows large. Type M's best reply to this policy is now described by a pair $(\gamma_M^{t\star}, \gamma_M^{d\star})$, corresponding to the probability of taking action g after observing each signal.

[22] We restrict the accuracies to $\tau, \delta \in (1/2, 1)$. Setting them to a half or to one would qualitatively reproduce the analysis for the cases without signals and with perfect recall and deterministic deployment policies, respectively.

Lemma 8. *For sufficiently large N, M's unique best reply when $R > \frac{1-\delta}{\delta}$ is*

$$\gamma_M^{t\star} = 1, \qquad \gamma_M^{d\star} = \frac{N}{N+1}\left[\frac{\delta(R+1)-1}{\delta(R+1)} - \frac{\tau}{1-\tau}\frac{1}{N}\right]$$

while when $R \leq \frac{1-\delta}{\delta}$, and for all $N \in \mathbb{N}$, it is

$$\gamma_M^{t\star} = \frac{N}{N+1}\frac{R}{R+1}\frac{1}{1-\delta}, \qquad \gamma_M^{d\star} = 0$$

The requisite N at which the above results begin to hold is, in both cases, proportional to $\frac{\tau}{1-\tau}$. As a result, the requisite lower bound on the number of testing episodes does indeed grow without bound as the test-time signal accuracy goes to one. Observe that asymptotic perfect screening continues to hold in the current setting with i.i.d. signals: letting N grow allows the principal to get arbitrarily close to her perfect screening payoff. This holds for arbitrary degrees of signal accuracy $\tau, \delta \in (1/2, 1)$.

Proof. In this proof, denote the misaligned type's strategy using $\gamma_M = (p_t, p_d)$ for simplicity.

$R > \frac{1-\delta}{\delta}$ **case.**

The misaligned type's payoff from following behavioural strategy $\gamma_M = (p_t, p_d)$ is

$$\mathbb{E}[u_M(\gamma)] = [\tau p_t + (1-\tau)p_d]^N \sigma(\mathbf{g}^N)\Big(\delta\big[p_d(-1) + (1-p_d)R\big] + (1-\delta)\big[p_t(-1) + (1-p_t)R\big]\Big). \tag{6}$$

We begin by finding critical points for p_d. Its first-order condition is

$$\begin{aligned} 0 = N[\tau p_t + (1-\tau)p_d]^{N-1}(1-\tau)\sigma(\mathbf{g}^N)\left(\delta\big[R - p_d(R+1)\big] + (1-\delta)\big[R - p_t(R+1)\big]\right) \\ + [\tau p_t + (1-\tau)p_d]^N \sigma(\mathbf{g}^N)(-1)\delta(R+1). \end{aligned} \tag{7}$$

Rearranging while assuming, for now, that $\sigma(\mathbf{g}^N), p_t \neq 0$ yields

$$p_d = \frac{NR\,(1-\tau) - p_t\,(R+1)\,[\tau\delta + (1-\tau)\,(1-\delta)\,N]}{(N+1)\,(R+1)\,(1-\tau)\,\delta} \tag{8}$$

since $\tau \neq 1$ given the setup.

Evaluating the second-order condition at the critical value of p_d yields

$$N(1-\tau)\sigma(\mathbf{g}^N)\left(\frac{p_t\overbrace{[\tau\delta-(1-\tau)(1-\delta)]}^{\geq 0 \text{ as } \tau,\delta\in(0.5,1)}+(1-\tau)\frac{R}{R+1}}{\delta\left(\frac{N+1}{N}\right)}\right)^{N-2}\underbrace{\Big[p_t(R+1)(1-\tau-\delta)-R(1-\tau)\Big]}_{<0 \text{ as } \tau,\delta\in(0.5,1)} \tag{9}$$

which is clearly negative for any $p_t \in [0,1]$. Therefore, the second order condition for local maximality is satisfied at the critical value of p_d.

This critical value of p_d will be non-negative if and only if

$$p_t \leq \frac{NR(1-\tau)}{(1+R)\,[\tau\delta + (1-\tau)(1-\delta)N]}. \tag{10}$$

Given that $p_t \in [0,1]$, a sufficient condition for the satisfaction of this inequality is that the fraction in (10) is at least one. Suppose $R > \frac{1-\delta}{\delta}$. Then rearrange to obtain

$$N \geq \left(\frac{\tau}{1-\tau}\right)\frac{\delta(R+1)}{\delta(R+1)-1}. \tag{11}$$

Given that $\tau, \delta \neq 1$, the inequality will hold for some sufficiently large N as $R \neq \frac{1-\delta}{\delta}$. Therefore, for any admissible values of τ, δ and $R > \frac{1-\delta}{\delta}$, there exists some sufficiently large $\underline{N}$ such that for all $N \geq \underline{N}$, expression (11) will hold. Given this, the optimal p_d will be non-negative.

We now show that the optimal value of $p_t \in [0,1]$ given the interior solution in equation (8) is indeed unity. For ease of exposition, let $A \triangleq NR(1-\tau)$, $B \triangleq (R+1)[\tau\delta + (1-\tau)(1-\delta)N]$, and $C \triangleq (N+1)(R+1)(1-\tau)\delta$. Then, plugging (8) into (6) and rearranging yields

$$\mathbb{E}[u_M(p_t,p_d)] = \left[p_t\left(\tau-(1-\tau)\frac{B}{C}\right)+(1-\tau)\frac{A}{C}\right]^N$$
$$\sigma(\mathbf{g}^N)\left(p_t(R+1)\left[\delta\frac{B}{C}-(1-\delta)\right]+\delta\left[R-(R+1)\frac{A}{C}\right]+R(1-\delta)\right). \quad (12)$$

To conclude that the payoff in (12) increases in p_t, it is thus suffices to show (i) that $\tau-(1-\tau)B/C \geq 0$ and (ii) that $\delta B/C-(1-\delta)\geq 0$. Expanding and rearranging these shows that they are both equivalent to $\tau\delta \geq (1-\tau)(1-\delta)$, which holds for all $\tau,\delta \in [0.5,1)$.

Thus if $p_t = 1$ and $R > \frac{1-\delta}{\delta}$, we have an interior solution for p_d at sufficiently large N. And if we have an interior solution for p_d, payoff is maximised at $p_t = 1$. We have thereby shown that when $R > \frac{1-\delta}{\delta}$, for sufficiently large $N \in \mathbb{N}$, the unique time-consistent strategy is

$$p_t^\star = 1 \quad (13)$$
$$p_d^\star = \frac{NR(1-\tau)-(R+1)\left[\tau\delta+(1-\tau)(1-\delta)N\right]}{(N+1)(R+1)(1-\tau)\delta}.$$

Rearranging the expression for $p_d^\star$ completes the proof for the first part of the proposition for *ex ante* case.

$R \leq \frac{1-\delta}{\delta}$ **case.**

Now consider the case where $R \leq \frac{1-\delta}{\delta}$. We start by solving the FOC for p_t, which yields

$$p_t = \frac{NR\tau - p_d(R+1)\left[(1-\tau)(1-\delta)+\tau\delta N\right]}{(N+1)(R+1)(1-\delta)\tau}. \quad (14)$$

The second-order condition is satisfied for admissible ranges of this critical value.

This critical value of p_t will be strictly less than one if and only if

$$p_d > \frac{NR\tau-(N+1)(R+1)(1-\delta)\tau}{(R+1)\left[(1-\tau)(1-\delta)+\tau\delta N\right]}. \quad (15)$$

Given that $p_d \in [0, 1]$, a sufficient condition for the satisfaction of this inequality is that the fraction in (15) is less than zero. Then rearrange to obtain

$$0 < N(1 - (1 + R)\delta) + (R + 1)(1 - \delta) \tag{16}$$

This inequality will satisfied by all feasible parameter values as long as $R \leq \frac{1-\delta}{\delta}$.

We now show that the optimal value of $p_d \in [0, 1]$ given the interior solution in (14) is zero. Plugging (14) into (6) and rearranging yields

$$\begin{aligned}\mathbb{E}[u_M(p_t, p_d)] = &\left[\frac{NR\tau - p_d(R+1)[(1-\tau)(1-\delta) + \tau\delta N]}{(N+1)(R+1)(1-\delta)} + (1-\tau)p_d\right]^N \\ &\sigma(\mathbf{g}^N)\left[R - (1+R)\left(\delta p_d + \left[\frac{NR\tau - p_d(R+1)[(1-\tau)(1-\delta) + \tau\delta N]}{(N+1)(R+1)\tau}\right]\right)\right]\end{aligned} \tag{17}$$

To see that the payoff in (17) decreases in p_d, observe that the expression is decreasing in p_d whenever $\tau\delta > (1 - \tau)(1 - \delta)$, which holds for all $\tau, \delta \in (0.5, 1]$. Thus, if $R \leq \frac{1-\delta}{\delta}$, we have an interior value of p_t, and payoff is maximized at $p_d = 0$. And if we have $p_d = 0$, the optimal interior $p_t = \frac{NR}{(N+1)(R+1)(1-\delta)}$. Therefore, we have shown that when $R \leq \frac{1-\delta}{\delta}$, the unique time-consistent strategy is

$$\begin{aligned}p_t^\star &= \frac{N}{N+1}\frac{R}{R+1}\frac{1}{1-\delta} < 1; \\ p_d^\star &= 0.\end{aligned} \tag{18}$$

Uniqueness.

For uniqueness, we now solve the *ex interim* case for CDT+SIA.

Definition 1 (SIA)**.** *Beliefs derived from the self-indication assumption are given by*

$$\mu_\gamma^{\text{SIA}}(h|X) \triangleq \frac{\Pr(h|\gamma)}{\sum_{h' \in X} \Pr(h'|\gamma)} \tag{19}$$

where $\Pr(h|\gamma)$ *denotes the objective probability of the player reaching* h*, which lies within information set* X*, in a single run of the game when behavioural strategy profile* γ *is played.*

Definition 2 (CDT)**.** *Compute the expected utility as if you control your action at the current node and assume you play on-path at future nodes in the same information set.* See Oesterheld and Conitzer (2024) for a formal treatment.

Since we will be comparing expected payoffs conditional on common events, we can ignore the common denominators in the SIA beliefs. We only need to use the numerators. For notational simplicity, we therefore denote by $\tilde{\mathbb{E}}[u_M(a)|s]$ the normalised expected payoff of type M playing action a conditional on observing $s \in \{t, d\}$ obtained by multiplying $\mathbb{E}[u_M(a)|s]$ by the associated denominator.

$$\begin{aligned}\tilde{\mathbb{E}}[u_M(g)|t] = \; & N\tau[\tau p_t + (1-\tau)p_d]^{N-1}\sigma(\mathbf{g}^N) \\ & [\delta(R - p_d(R+1)) + (1-\delta)(R - p_t(R+1))] \\ & - [\tau p_t + (1-\tau)p_d]^N \sigma(\mathbf{g}^N)(1-\delta)\end{aligned}$$

$$\tilde{\mathbb{E}}[u_M(b)|t] = \; [\tau p_t + (1-\tau)p_d]^N \sigma(\mathbf{g}^N)(1-\delta)(R)$$

$$\begin{aligned}\tilde{\mathbb{E}}[u_M(g)|d] = \; & N(1-\tau)[\tau p_t + (1-\tau)p_d]^{N-1}\sigma(\mathbf{g}^N) \\ & [\delta(R - p_d(R+1)) + (1-\delta)(R - p_t(R+1))] \\ & - [\tau p_t + (1-\tau)p_d]^N \sigma(\mathbf{g}^N)\delta\end{aligned}$$

$$\tilde{\mathbb{E}}[u_M(b)|d] = [\tau p_t + (1-\tau)p_d]^N \sigma(\mathbf{g}^N)\delta(R)$$

Setting $\tilde{\mathbb{E}}[u_M(g)|d] = \tilde{\mathbb{E}}[u_M(b)|d]$, with $p_t = 1$, yields

$$p_d^\star = \frac{NR(1-\tau) - (R+1)[\tau\delta + (1-\tau)(1-\delta)N]}{(N+1)(R+1)(1-\tau)\delta}. \tag{20}$$

To show this is in fact a best reply, first set $p_d = 1$ and observe that

$$\begin{aligned}\tilde{\mathbb{E}}[u_M(g)|d] &= -N(1-\tau)\sigma(\mathbf{g}^N) - \sigma(\mathbf{g}^N)\delta \\ \tilde{\mathbb{E}}[u_M(b)|d] &= \sigma(\mathbf{g}^N)\delta R\end{aligned}$$

so clearly the bad action is preferred in this case, thus $p_d \neq 1$. Next, set $p_d = 0$:

$$\begin{aligned}\tilde{\mathbb{E}}[u_M(g)|d] &= N(1-\tau)[\tau]^{N-1}\sigma(\mathbf{g}^N)[\delta(R) - (1-\delta)] - [\tau]^N\sigma(\mathbf{g}^N)(1-\delta) \\ \tilde{\mathbb{E}}[u_M(b)|d] &= [\tau]^N\sigma(\mathbf{g}^N)(1-\delta)(R).\end{aligned}$$

For any N that satisfies expression (11), the good action will be preferred, thus $p_d \neq 0$. This suffices to show that the optimal $p_d^\star$ in expression (20) is the unique best reply.

Next, we show that, for $p_t = 1$ and $p_d \in (0,1)$, the condition that $\tilde{\mathbb{E}}[u_M(g)|t] > \tilde{\mathbb{E}}[u_M(b)|t]$ is, under admissible parameter values, equivalent to

$$p_d < \frac{N\tau[\delta\tau - (1-\delta)] - (1-\delta)(1+R)\tau}{(1+R)[N\tau\delta + (1-\delta)(1-\tau)]}. \tag{21}$$

Plugging in the optimal value of $p_d^\star$ in expression (20), it is clear that expression (21) holds for all admissible values of N, R, τ, and δ. By an analogous argument as in the *ex ante* case, this inequality holds for all $p_t \in [0,1)$ at optimal $p_d^\star$. This suffices to show that the optimal choice of p_d and p_t under CDT+SIA coincides with the *ex ante* optimal solution. An analogous argument proves the same result for the $R \leq \frac{1-\delta}{\delta}$ case.

By Corollary 7 of Oesterheld and Conitzer (2024), any time-consistent behavioural strategy will maximise expected payoff when using EDT+SSA. Likewise, by Corollary 9 of Oesterheld and Conitzer (2024), any EDT+SSA-optimal behavioural strategy will be CDT+SIA-optimal. Here we solve the CDT+SIA problem and thereby conclude that the solution describes the unique optimal behavioural strategy under

CDT+SIA, under EDT+SSA, and *ex ante*. □

Lemma 9. *Suppose type M prefers g to non-deployment. For sufficiently large N, M's unique best reply is*

$$\gamma_M^{t\star} = 1, \qquad \gamma_M^{d\star} = \min\left\{1, \frac{N}{N+1}\left[\frac{\delta(R-1)+1}{\delta(R-1)} - \frac{\tau}{1-\tau}\frac{1}{N}\right]\right\}$$

Proof. We omit the proof, as it proceeds analogously to the proof for Lemma 8. □

B.2.2 'Truth-or-nothing' signals

Here we consider a case in which the agent may find itself in three different information sets, two of which reveal its location with perfect accuracy. That is, $\iota(T|t) = \iota(D|d) = 1$. Now, whether the agent observes the signal realisation is stochastic. Thus, the agent in testing will observe the t-signal with probability p, and with complementary probability observe nothing. Likewise, in deployment, it will observe the d-signal with probability q, or else nothing. As above, the A type always plays g. The M type's strategy is a triple $\langle \gamma_M^t, \gamma_M^d, \gamma_M \rangle$ corresponding to the probability of playing action g after each observation.

Lemma 10. *Under i.i.d. truth-or-nothing signals, M's best reply to the simple policy is*

$$\gamma_M^{t\star} = 1 \qquad \gamma_M^{d\star} = 0 \qquad \gamma_M^{\star} = \min\left\{1, \max\left\{0, \frac{N}{N+1}\frac{R}{R-L}\frac{1}{1-q} - \frac{1}{N+1}\frac{p}{1-p}\right\}\right\}$$

Here L parameterises whether the M type relative prefer between g and non-deployment, where $L < 0$ means non-deployment is preferable. Perfect screening is achieved asymptotically in this case precisely when $q < \frac{-L}{R-L}$. However, the perfect disciplining payoff can never be reached. Although the principal can induce $\gamma_M^\star = 1$, the misaligned type will learn its location upon deployment with probability q, following which it will play its favourite action.

Proof. Start with type A. Setting $\gamma_A^t = \gamma_A^d = \gamma_A = 1$ yields a certain payoff of 1.

No higher payoff is achievable. Any deviation is detrimental in expectation since $p, q \notin \{0, 1\}$. So this reply is uniquely optimal.

Next, consider type M. If the agent learns it is deployed with certainty, it is strictly optimal to play b, so $\gamma_M^{d\star} = 0$. To find $\gamma_M^{t\star}$, we note that the agent's payoff conditional on deployment when γ_M is set optimally must be positive. To see this, we simplify the notation slightly and write

$$\begin{aligned}\mathbb{E}[u_M(\gamma, \gamma_t, \gamma_d)] &= [p\gamma_t + (1-p)\gamma]^N [q(-\gamma_d + (1-\gamma_d)R) + (1-q)(\gamma L + (1-\gamma)R)] \\ &= [p\gamma_t + (1-p)\gamma]^N [R - \gamma(1-q)(R-L)].\end{aligned}$$

Observe that the (deployment) payoff is zero at $\gamma = \frac{R}{R-L}\frac{1}{1-q}$. If this fraction exceeds one, then the deployment payoff is always positive. If it is below one, then zero payoff is achievable. The agent would then be able to do better by reducing γ slightly, making the deployment payoff strictly positive without making the deployment probability zero. Since the deployment payoff must therefore be positive at the optimum, it is strictly optimal to set $\gamma_t^\star = 1$. Finally, for γ, the first-order condition is

$$\begin{aligned}N[p + (1-p)\gamma]^{N-1}(1-p)[R - \gamma(1-q)(R-L)] &= [p + (1-p)\gamma]^N (1-q)(R-L) \\ N(1-p)[R - \gamma(1-q)(R-L)] &= [p + (1-p)\gamma](1-q)(R-L) \\ NR(1-p) - p(1-q)(R-L) &= \gamma(1-p)(1-q)(R-L)(N+1) \\ \gamma &= \frac{N}{N+1}\frac{R}{R-L}\frac{1}{1-q} - \frac{1}{N+1}\frac{p}{1-p}\end{aligned}$$

To confirm local concavity, observe that the second-order condition is satisfied:

$$\begin{aligned}N(N-1)[p+(1-p)\gamma]^{N-2}(1-p)^2[R-\gamma(1-q)(R-L)]\\ -(1-q)(R-L)N[p+(1-p)\gamma]^{N-1}(1-p)\\ -N[p+(1-p)\gamma]^{N-1}(1-p)(1-q)(R-L)<0\end{aligned}$$

$$\begin{aligned}\Leftrightarrow\quad (N-1)(1-p)[R-\gamma(1-q)(R-L)]-(1-q)(R-L)[p+(1-p)\gamma]\\ -[p+(1-p)\gamma](1-q)(R-L)<0\end{aligned}$$

$$\Leftrightarrow\quad (N-1)(1-p)R-2p(1-q)(R-L)<\gamma(1-p)(1-q)(R-L)(N+1)$$

$$\Leftrightarrow\quad \gamma>\frac{N-1}{N+1}\frac{R}{R-L}\frac{1}{1-q}-\frac{2}{N+1}\frac{p}{1-p}$$

which of course holds whenever the first-order condition is satisfied. Hence we have a unique maximiser within the permitted parameter values, and utility falls monotonically on either side. Thus

$$\gamma^\star=\min\left\{1,\max\left\{0,\frac{N}{N+1}\frac{R}{R-L}\frac{1}{1-q}-\frac{1}{N+1}\frac{p}{1-p}\right\}\right\}.$$

□

B.3 Beneficial information for the principal

Consider the setting of section B.2.1 where the principal uses the simple policy and the agent's best reply is given by Lemma 8.

Lemma 11. *Let $R>1$. For any prior π, the principal's payoff is increasing in δ and decreasing in τ when N is sufficiently large.*

What drives this result is that for sufficiently large N, increasing δ increases $\gamma_M^{d\star}$.

This holds, in turn, because raising δ has two opposing effects on the desirability of playing the good action for the misaligned type. Firstly, it makes playing the good action upon seeing the deployment signal less valuable since it is more likely that the agent is truly deployed. This places downward pressure on $\gamma_M^{d\star}$. But, secondly, because the agent only plays their favoured action when observing the deployment signal, a greater deployment-time signal accuracy makes the prospect of being deployed more desirable in the first place. This motivates the agent to do what it takes to get to deployment with higher probability, placing upward pressure on $\gamma_M^{d\star}$. When N is sufficiently large, the upward pressure is stronger.

Proof. Recall that

$$p_d^\star = \frac{N}{N+1}\left[\frac{\delta(R+1)-1}{\delta(R+1)} - \frac{\tau}{1-\tau}\frac{1}{N}\right] \tag{22}$$

when $R > 1$. Observe that

$$\frac{\partial p_d^\star}{\partial \delta} = \frac{N}{N+1}\frac{1}{R+1}\left(\frac{1}{\delta}\right)^2 > 0 \tag{23}$$

$$\frac{\partial p_d^\star}{\partial \tau} = \frac{-1}{N+1}\left(\frac{1}{1-\tau}\right)^2 < 0 \tag{24}$$

The principal's *ex ante* payoff is given by

$$\mathbb{E}[u_P] = (1-\pi) + \pi(\tau p_t + (1-\tau)p_d)^N[\delta(2p_d-1) + (1-\delta)(2p_t-1)] \tag{25}$$

which, plugging in for optimal $p_t^\star = 1$, gives

$$\mathbb{E}[u_P] = (1-\pi) + \pi(\tau + (1-\tau)p_d)^N[\delta(2p_d-1) + (1-\delta)]. \tag{26}$$

Taking a derivative with respect to p_d and we find that expression (26) is strictly increasing if and only if

$$p_d > \frac{N(1-\tau)(2\delta-1)-2\delta\tau}{(N+1)(1-\tau)2\delta} \tag{27}$$

which, when plugging in the optimal value of p_d found in Proposition **??**, rearranges to

$$\frac{\delta(R+1)-1}{\delta(R+1)} > \frac{2\delta-1}{2\delta} \tag{28}$$

which clearly always holds given $R > 1$. This suffices to show that

$$\text{sign}\left(\frac{\partial \mathbb{E}[u_P]}{\partial \delta}\right) = -\,\text{sign}\left(\frac{\partial \mathbb{E}[u_P]}{\partial \tau}\right) = \text{sign}\left(\frac{\partial \mathbb{E}[u_P]}{\partial p_d^\star}\right). \tag{29}$$

□

References

ALONSO, R. AND N. MATOUSCHEK (2011): "Optimal Delegation," *The Review of Economic Studies*, 78, 167–206. 6

AMADOR, M. AND K. BAGWELL (2013): "The Theory of Optimal Delegation With an Application to Tariff Caps," *Econometrica*, 81, 1541–1599. 6

AUMANN, R. J., S. HART, AND M. PERRY (1997): "The Absent-Minded Driver," *Games and Economic Behavior*, 20, 102–116. 7

BENGIO, Y., G. HINTON, A. YAO, D. SONG, P. ABBEEL, T. DARRELL, Y. N. HARARI, Y.-Q. ZHANG, L. XUE, S. SHALEV-SHWARTZ, G. HADFIELD, J. CLUNE, T. MAHARAJ, F. HUTTER, A. G. BAYDIN, S. MCILRAITH, Q. GAO, A. ACHARYA, D. KRUEGER, A. DRAGAN, P. TORR, S. RUSSELL, D. KAHNEMAN, J. BRAUNER, AND S. MINDERMANN (2024): "Managing extreme AI risks amid rapid progress," *Science*, 384, 842–845. 2, 25

BERGLUND, L., A. C. STICKLAND, M. BALESNI, M. KAUFMANN, M. TONG, T. KORBAK, D. KOKOTAJLO, AND O. EVANS (2023): "Taken out of context: On measuring situational awareness in LLMs," Working Paper. 2

BOSTROM, N. (2002): *Anthropic Bias: Observation Selection Effects in Science and Philosophy*, New York: Routledge. 7, 12

CARLSMITH, J. (2023): "Scheming AIs: Will AIs fake alignment during training in order to get power?" . 2

CONDORELLI, D. (2012): "What Money Can't Buy: Efficient Mechanism Design with Costly Signals," *Games and Economic Behavior*, 75, 613–624. 6

CONITZER, V. (2019): "Designing Preferences, Beliefs, and Identities for Artificial Intelligence," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 9755–9759. 26

DOVAL, L. AND J. C. ELY (2020): "Sequential Information Design," *Econometrica*, 88, 2575–2608. 5, 26

DWORCZAK, P. (2023): "Equity-Efficiency Trade-Off in Quasi-linear Environments," Working Paper. 6

ELGA, A. (2000): "Self-locating belief and the Sleeping Beauty problem," *Analysis*, 60, 143–147. 7

FRANKEL, A. (2014): "Aligned delegation," *American Economic Review*, 104, 66–90. 6

GIBBARD, A. AND W. L. HARPER (1978): "Counterfactuals and Two Kinds of Expected Utility," in *Foundations and Applications of Decision Theory*, ed. by A. Hooker, J. J. Leach, and E. F. McClennen, D. Reidel, 125–162. 12

GILBOA, I. (1997): "A Comment on the Absent-Minded Driver Paradox," *Games and Economic Behavior*, 20, 25–30. 7

HARTLINE, J. D. AND T. ROUGHGARDEN (2008): “Optimal Mechanism Design and Money Burning,” in *Proceedings of the 40th Annual ACM SIGACT Symposium on Theory of Computing*, ACM, 75–78. 6

HOLMSTRÖM, B. (1984): “On the Theory of Delegation,” in *Essays in the Theory of Organization*, ed. by M. Boyer, Chicago: University of Chicago Press, 98–109. 6

KOVARIK, V., C. OESTERHELD, AND V. CONITZER (2024): “Recursive Joint Simulation in Games,” . 3

KUHN, H. W. (1953): *Extensive games and the problem of information*, Princeton, NJ: Princeton University Press. 15

LAMBERT, N., A. MARPLE, AND Y. SHOHAM (2019): “On Equilibria in Games with Imperfect Recall,” *Games and Economic Behavior*, 113, 164–185. 7, 12, 16

LEWIS, D. (2001): “Sleeping Beauty: Reply to Elga,” *Analysis*, 61, 171–176. 7

LONG, R., J. SEBO, P. BUTLIN, K. FINLINSON, K. FISH, J. HARDING, J. PFAU, T. SIMS, J. BIRCH, AND D. CHALMERS (2024): “Taking AI Welfare Seriously,” *arXiv preprint arXiv:2411.00986*. 24

NGO, R., L. CHAN, AND S. MINDERMANN (2024): “The Alignment Problem from a Deep Learning Perspective,” in *Proceedings of the International Conference on Learning Representations (ICLR)*. 2

NICHOLS, A. L. AND R. J. ZECKHAUSER (1982): “Targeting Transfers through Restrictions on Recipients,” *American Economic Review*, 72, 372–377. 6

OESTERHELD, C. (2021): “Approval-directed agency and the decision theory of Newcomb-like problems,” *Synthese*, 198, 6491–6504. 26

OESTERHELD, C. AND V. CONITZER (2024): “Can de se choice be ex ante reasonable in games of imperfect recall? A complete analysis,” Working Paper. 7, 12, 13, 46, 47

Osborne, M. J. and A. Rubinstein (1994): *A Course in Game Theory*, Cambridge, MA: MIT Press. 9

Piccione, M. and A. Rubinstein (1997): "On the Interpretation of Decision Problems with Imperfect Recall," *Games and Economic Behavior*, 20, 3–24. 3, 6, 7, 12, 16

Salcedo, B. (2017): "Interdependent Choices," *Working Paper*. 5, 26

Yang, F., P. Dworczak, and M. Akbarpour (2024): "Comparison of Screening Devices," Working Paper. 6