Narrow Inference and Incentive Design[*]

Alexander Clyde[†]

August 20, 2025

[CLICK HERE FOR LATEST VERSION]

Abstract

There is evidence that people struggle to do causal inference in complex multidimensional environments. This paper explores the consequences of this in a principal-agent setting. A principal chooses a mechanism to screen an agent. The agent makes choices on multiple dimensions, and infers the effect of each action separately without properly controlling for the other actions. I fully characterize the principal's optimal mechanism when facing an agent who does such 'narrow' inference, and contrast it with their optimal mechanism when the agent is fully rational. I demonstrate when the principal can exploit narrow inference and in what cases they lose out.

Keywords: Behavioural Mechanism Design, Screening, Narrow Bracketing, Misspecified Models.

JEL Classification: D90, D02, D82

[*] I am grateful to my advisors Ran Spiegler and Philippe Jehiel, for their support and guidance at all stages of this project. I also thank Martin Cripps, Nathan Hancart, Deniz Kattwinkel, Vasiliki Skreta, Duarte Gonçalves, Michael Thaler, Mark Armstrong, Nikita Roketskiy, Elena Ashtari Tafti, Dajana Xhani and audiences at Humboldt-Universität zu Berlin, Bilkent University, Edinburgh University, the Midwest Theory Conference at Virginia Tech, the Barcelona Summer Forum Workshop on Preferences, Bounded Rationality, and Strategic Interactions, the Conference on Economic Design at University of Essex, the Nordic Economic Theory Workshop at University of Copenhagen, the SAET Conference in Ischia and the Stony Brook International Game Theory Conference for helpful questions and comments. I gratefully acknowledge financial support from an Economic and Social Research Council studentship (1923039) and from Leverhulme Trust grant RPG-2023-120.

[†] Department of Economics, Aalto University School of Business; Address: Ekonominaukio 1, FI-02150 Espoo, Finland; Email: alexander.clyde.econ@gmail.com or alexander.clyde@aalto.fi

1 Introduction

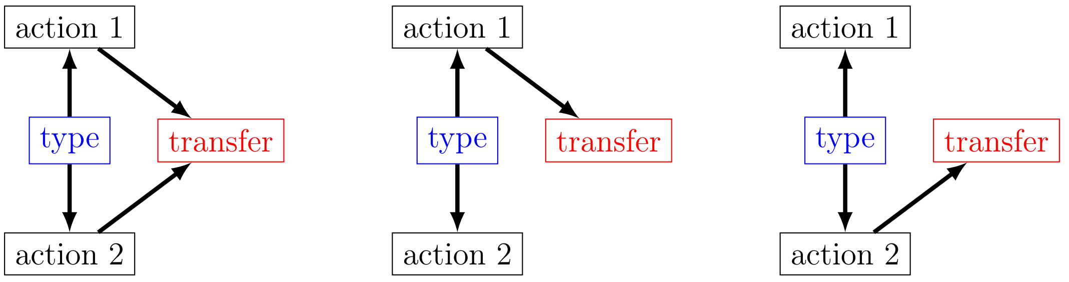
Understanding the incentives we face requires understanding what the causal consequences of our actions are on outcomes that we care about. For example, workers have to form beliefs about how their choices of effort, occupation and education affect their ultimate earnings. Economic models often implicitly assume that people can form these beliefs by making sophisticated causal inference from any available data, and that they consider the many choices they make jointly as if they were one over-arching decision problem. However, work in experimental economics suggests that people both fail to consider their choices jointly and to adjust correctly for correlation in the data they are using for belief formation.[1]

Since causal understanding matters for the perception of incentives, when designing incentives it is important to take into account people's limited ability to perform causal inference. In this paper, I provide a full characterization of how a principal should design incentives for an agent who forms beliefs about the effects of their actions according to a procedure I call *narrow inference*. This model of belief formation is able to link models of causal misperceptions from the literature on misspecified models with work in behavioural economics on narrow choice bracketing.

To understand narrow inference, consider a firm that is determining its wage structure. A worker in the firm has to decide whether to make human capital investments in technical skills and/or managerial skills. Under narrow inference, the worker estimates the earnings benefits of acquiring each skill narrowly and separately, by comparing the average wage of workers in the firm who have the skill in question with the average wage of workers without that skill. In forming beliefs about the effect of any individual action in this way, they fail to control for other dimensions of action. This leads to a confounding bias that distorts the worker's perception of the size of earnings benefits from obtaining different forms of human capital. Figure 1 illustrates this bias using the Directed Acyclic Graph (DAG) notation of Pearl (2009). The extent of this misperception affects how the firm wants to design their wage structure.

In this paper I analyze such incentive design problems. I explore how a principal would design an incentive mechanism if they knew the agent had this form of bounded rationality. I obtain a result characterizing the principal's optimal mech-

[1] In the literature review I discuss work on 'narrow bracketing' from behavioural economics, and work in experimental economics on correlation neglect and causal misperceptions.



(a) True causal model (b) Dimension 1 beliefs (c) Dimension 2 beliefs

Figure 1: DAG illustrating confounding bias with two dimensions of action.

anism with an agent who makes narrow inference as the solution to a zero-sum game. I use this result to demonstrate in what cases the principal benefits from the agent making narrow rather than fully rational inference, and in what cases the principal does not. I then explore the effect of splitting the costs and benefits of the actions across more dimensions. In doing so, I demonstrate that under narrow inference the splitting and merging of dimensions becomes a quantitatively significant margin of incentive design.

In the screening model, the agent faces a binary decision problem on whether to take an action or not on multiple dimensions. The principal chooses a function mapping the agent's actions to a transfer. There is a large population of agents who differ according to a single dimensional type variable that affects the predictable costs and benefits of the actions. The utility of the agents can be decomposed into a type-dependent predictable part that they fully understand, and a transfer part that they need to infer. The principal's choice of transfer function screens different types of agents into choosing different actions. In the absence of bounded rationality — and if the agent therefore knows and understands the true transfer function— this is a standard monopolistic screening problem.

An agent who makes narrow inference calculates the effect of each dimension's action on the transfer separately. Their beliefs about the effect of an action on a given dimension must be consistent with data on the population level average transfer conditional on that action. The difference between the average population level transfers between any two actions in a given dimension is then used to estimate the relative effect of each action on the transfer. This is a naive way to estimate the 'treatment effect' of any action. It can lead to distorted beliefs if the distribution over actions are correlated across dimensions, something that is possible due to joint dependence on the type variable.

In what follows, I develop the firm-worker application to illustrate narrow inference and to demonstrate how the principal might benefit from agents using narrow inference.

Example 1. The agent has two dimensions of action, whether to obtain technical skills $a_1 = 1$ or not $a_1 = 0$ and whether to obtain managerial skills $a_2 = 1$ or not $a_2 = 0$. The principal, a firm, sets an earnings schedule for roles within the firm that depends jointly on these two actions $t : A_1 \times A_2 \to \mathbb{R}$. There are three types of agent $s \in \{0, 1, 2\}$. The probabilities of the types are denoted p_0, p_1, $p_2 \in (0, 1)$ respectively. The agent's utility depends on their type s, their actions a_1, a_2 and the transfer t.

$$t(a_1, a_2) - (3 - s)(a_1 + a_2)$$

Where $-(3 - s)(a_1 + a_2)$ is the predictable utility cost of making human capital investments for the worker. Suppose that the principal wants to implement that the type $s = 0$ chooses neither action, the type $s = 1$ obtains technical skills $a_1 = 1$ but not managerial skills $a_2 = 0$, while the highest type $s = 2$ obtains both $a_1 = a_2 = 1$. An earnings function that ensures that rational agents will act this way must satisfy the following incentive constraints. The first ensures that type $s = 1$ chooses $(1, 0)$ over $(0, 0)$ and the second ensures type $s = 2$ chooses action $(1, 1)$ over $(1, 0)$.

$$t(1, 0) - 2 \geq t(0, 0)$$
$$t(1, 1) - 2 \geq t(1, 0) - 1$$

Choosing t such that these incentive constraints bind allows the principal to minimize the earnings paid to types $s = 1$ and $s = 2$. This means $t(1, 1) > t(1, 0) > t(0, 0)$. These local incentive constraints binding suffices for all incentive constraints to hold.

Now consider what happens if the agent uses narrow inference, but the transfer function is fixed at those that bind for rational incentive compatibility. They expect the earnings from any action to be the average population level earnings of those who have taken that action. For an agent of type $s = 1$ making narrow inference to want to invest in technical skills requires that

$$\frac{p_1}{p_1 + p_2} t(1, 0) + \frac{p_2}{p_1 + p_2} t(1, 1) - 2 \geq t(0, 0)$$

Their narrow beliefs about the earnings from acquiring technical skills pool the earnings of both the types who obtain these skills, who have proportion $p_1 + p_2$ in the population of types. The narrow beliefs about the earnings from not acquiring technical skills are in line with rational beliefs, since only the lowest type $s = 0$ doesn't acquire these skills. Since $t(1,1) > t(1,0)$, the narrow perception of the expected earnings benefit of technical skills is biased upward from the true effect. It fails to adjust for the fact that a type $s = 1$ agent who is on the margin between obtaining technical skills does not obtain managerial skills and thus has lower future earnings than the average worker who obtains technical skills.

Similarly for the type $s = 2$ agent to want to obtain managerial skills under narrow inference requires that

$$t(1,1) - 1 \geq \frac{p_1}{p_1 + p_0} t(1,0) + \frac{p_0}{p_1 + p_0} t(0,0)$$

As $t(1,0) > t(0,0)$, their perception of the earnings from not obtaining managerial skills is biased downwards. The expected earnings for those who do not obtain managerial skills mixes the earnings of those who get technical skills and those who do not. It is therefore less than the earnings obtained by the types on the margin of obtaining managerial skills, type $s = 2$, all of whom obtain technical skills.

This upward bias in the incentives the agent perceives allows the principal to implement the same action choices for each type while reducing earnings across the type distribution. △

My analysis of the design problem proceeds as follows. First, in order to contrast the principal's optimal mechanism when agents make narrow inference to that with fully rational agents, I state results adapted from Carroll (2017) which describe the principal's optimal mechanism for the rational benchmark. Under a standard regularity assumption, the principal's optimal mechanism is fully separable across dimensions. Facing such a mechanism, the agent chooses a strategy that on each dimension selects the action if and only if their type is above a dimension-specific threshold.

I then obtain a result characterizing the principal's optimal mechanism under narrow inference. In this characterization, the principal plays a zero-sum game against an adversarial player who can shrink the agent's predictable utility from the actions on any dimension by some constant factor that lies between one and zero, with the shrinkage factors summing to one across dimensions. The principal's optimal mechanism under narrow inference then solves the same problem as in the rational benchmark except with the shrunk predictable utilities. The result allows us to both solve for the principal's optimal mechanism for specific parameterizations, and also enables us to easily make comparisons with the principal's optimal mechanism for rational agents.

I show using the characterization result what effect narrow inference has on this implemented threshold strategy and on the principal's welfare relative to the rational benchmark. This involves considering two different cases; one where the agent's actions have predictable utility costs to the agent but benefits to principal on all dimensions, and one where all actions have predictable benefits to the agents but costs to the principal.[2] The first case fits the firm-worker application, where human capital investments have acquisition costs for the worker but increase profitability for the firm. For an application that fits the second case, consider a buyer-seller setting. The principal is a seller of two goods, such as computers and software, for which the ultimate price can be opaque to the buyer and can be affected by both purchases in a joint way. The buyer gets positive predictable utility from buying a more advanced computer and from better software. They make narrow inference about what the effect of each choice is on the ultimate price they are paying for their computing.

In the case where actions predictably cost the agent but benefit the principal, when facing the principal's optimal mechanism the agent's thresholds are lower under narrow inference than when they are in the rational benchmark and the principal is always weakly better-off under narrow inference. Lower thresholds mean a greater proportion of types take the action on any dimension. On the other hand, when actions predictably cost the principal but benefit the agent then thresholds are higher and the principal is always weakly worse-off under narrow inference. These effects are the consequence of the agent overestimating the causal effects of their actions on transfers due to confounding bias. Taking the action on one dimension is associated with taking the action on others, as both are chosen by higher type agents, and when agents neglect this their beliefs about the effect of the actions are biased upwards. I show that the principal's gains and losses from

[2] Whether the actions have positive or negative predictable utility for the agent is not an arbitrary re-labelling of the actions, as we normalize the predictable utility from the zero action to zero. Without the normalization the result would be stated in terms of the difference in predictable utilities between the two actions.

narrow inference are purely due to this confounding, as under narrow inference the principal's optimal mechanism has a separable transfer function just as in the rational case.

I then consider what happens if total predictable utilities to the principal and agent are either split across or merged over action dimensions. I show that under narrow inference the principal benefits from splitting dimensions when actions are costly to the agent but benefit the principal, but wants to merge dimensions when actions benefit the agent but cost the principal. Since in the symmetric case splitting and merging dimensions has no effect in the rational benchmark, I then explore the effect of splitting over a growing number of symmetric action dimensions. If there are enough dimensions then in the principal's optimal mechanism under narrow inference either all types of agents take the actions on all dimensions, or no agents take the actions on any dimension. The principal can implement these strategies with transfers that shrink to zero as the number of dimensions grows. Thus, narrow inference has a particularly stark effect in the symmetric case.

Finally, I show that the principal can always achieve their optimum by choosing a transfer function that loads all incentives onto only one dimension of an additive transfer function. This is what is behind the result that with a growing number of symmetric dimensions, the transfer shrinks to zero. Under the one-dimension loaded transfer function, the agent over-counts the effect on the transfer that arises from taking the action on only one particular loaded dimension as also arising from taking the action on other dimensions. As the number of dimensions tends to infinity, the over-counting effect becomes larger and less and less actual transfer is required in order to incentivize the desired actions.

The paper takes a step towards understanding how errors in causal inference might affect how we should design economic incentives by studying a novel model of belief formation. In the Online Appendix, I show the robustness of the conclusions of this paper to an alternative participation constraint and also to dropping the regularity assumption on the type distribution that is made throughout the paper.

Literature Review

Experimental work in psychology and economics documents that people make inferential errors similar to narrow inference. Enke and Zimmermann (2019) find subjects fail to adjust for correlation between multiple information sources. Similar logic extends to predictive tasks, in He and Kučinskas (2024) subjects' forecasting performance deteriorates when information from a single variable is split into two. Fernbach et al. (2010) present evidence suggesting people focus narrowly on a few variables when trying to make causal predictions. In line with this, Graeber (2023) finds subjects ignore the effect of variables that are not directly involved in a predictive task despite these variables containing valuable information.

Narrow inference involves causal misperceptions, but also thinking about decision problems narrowly. The literature on narrow bracketing considers decision makers who break decision problems into smaller sub-problems without accounting for how these decisions interact in the larger joint problem. Work in this area: Tversky and Kahneman (1981), Thaler (1985), Thaler (1999), Read et al. (1999), Rabin and Weizsäcker (2009), has both documented evidence for and explored the theoretical implications of narrow decision making. Recent work exploring theoretical foundations for narrow behaviour includes Kőszegi and Matějka (2020), who use a model of costly information acquisition to explain both mental accounting and naive diversification. Lian (2021) builds a theory of 'narrow thinking', which models decision makers as playing an incomplete information bayesian game between multiple-selves. Camara (2022) shows that computability constraints imply a form of narrow choice bracketing. Vorjohann (2024) presents an axiomatization that can resolve tensions between narrow decision making and global budget balance.

In modelling agents with narrow causal perceptions, this paper builds on work studying decision making by agents using misspecified models of how action choices map into consequences. There is a growing literature on the Berk-Nash Equilibrium of Esponda and Pouzo (2016), a solution concept founded as the limit of a process of misspecified learning; (Heidhues et al., 2018; Frick et al., 2020; Bohren and Hauser, 2021; Fudenberg et al., 2021). Another connected literature is that on modelling causal misperceptions using Bayesian Networks; (Spiegler, 2016; Eliaz and Spiegler, 2020). Schumacher and Thysen (2022) use this Bayesian Network approach in a principal-agent moral hazard problem where the agent has causal misperceptions of how their actions map into output. In Eliaz and Spiegler (2024) a Bayesian Network formalism is used to model the design of narratives for misspecified news consumers by media organizations.

Earlier work on design when agents misperceive incentives by Rubinstein (1993) and Piccione and Rubinstein (2003) explores monopolistic pricing when customers have a coarse misperception of any pricing strategy. Eyster and Rabin (2005) consider a solution concept for games where players neglect correlation between information and opposing players' actions, and apply their concept to bilateral trade and auction settings. Jehiel (2005) develops an equilibrium concept for extensive form games —Analogy Based Expectation Equilibrium (ABEE)— in which players have coarse misperceptions of other players' strategies. The behaviour of the principal and the agent under narrow inference can be formulated as an ABEE of an extensive form game, and I discuss this in more detail in Online Appendix Section A. The first papers to explicitly apply the ABEE concept to design problems are Jehiel (2011) and Jehiel and Mierendorff (2024) who study auction design where bidders receive limited feedback. A paper explicitly studying narrow bracketing and auction design is Eisenhuth (2019), who uses a prospect theoretic formulation of narrow bracketing.

In contributing to the small literature on mechanism design where agents use misspecified models, this paper also contributes to a larger literature on mechanism design that takes into account agents' limited rationality in a variety of other dimensions. A detailed review can be found in Kőszegi (2014). This includes work in contract theory (Eliaz and Spiegler, 2006), (Heidhues and Kőszegi, 2010), (Herweg et al., 2010) and optimal taxation (O'Donoghue and Rabin, 2006), (Spinnewijn, 2015), (Farhi and Gabaix, 2020), (Lockwood, 2020).

2 Model

An agent faces a multidimensional decision. Let $A = \{0,1\}^n$ be the agent's set of feasible action profiles. I refer to $i \in \{1, \ldots, n\} \equiv N$ as a dimension, such that a_i is the agent's action in dimension i. The agent has a type that lies in a bounded interval $s \in S \equiv [0,1]$. This type is drawn from an atomless distribution that admits a density $p(s)$ such that $p(s) > 0$ for all $s \in S$. Denote the cdf of the distribution by $P(s) = \int_0^s p(\tilde{s})d\tilde{s}$.

The dimension i action a_i generates a predictable utility $v_i(s)a_i$, where $v_i(s)$ is strictly increasing, continuously differentiable in s and can be positive or negative. In addition to the predictable utility, the agent receives utility from a transfer t that needs to be inferred. The utility of an agent of type s, choosing action a with transfer $t \in \mathbb{R}$ is

$$u(s, a, t) = \sum_{i \in N} v_i(s) a_i + t \tag{1}$$

The principal derives benefit/costs $w_i a_i$ from an action a_i, where $w_i \in \mathbb{R}$. The transfer $t \in \mathbb{R}$ represents a zero-sum transfer of surplus between the agent and the principal. The principal's payoff given actions a and transfer $t \in \mathbb{R}$ is

$$W(a, t) = -t + \sum_{i \in N} w_i a_i \tag{2}$$

Throughout the paper, I make the following standard regularity assumption on the type distribution. In Section D of the Online Appendix, I explore the implications of relaxing this assumption.

Assumption 1. For every dimension $i \in N$

$$\phi_i(s) = v_i(s) - \frac{1 - P(s)}{p(s)} v_i'(s)$$

is strictly increasing in $s \in S$. We refer to this property as increasing virtual values (IVV).

2.1 Mechanisms

I focus on a natural class of indirect mechanisms. Before the agent takes any actions, the principal commits to a mechanism, which consists of a function mapping actions to transfers $t : A \to \mathbb{R}$. After learning their type, the agent chooses a distribution over actions according to a strategy $g : S \to \Delta(A)$. The marginal distribution over actions in dimension i is denoted by $g_i(a_i|s) = \sum_{a_{-i} \in A_{-i}} g(a_i, a_{-i}|s)$.

For a transfer function $t \in \mathbb{R}^A$, given a strategy g the expected payoff for the principal is

$$W(t, g) = \int_0^1 \sum_{a \in A} [-t(a) + \sum_{i \in N} w_i a_i] g(a|s)\, p(s) ds \tag{3}$$

The restriction to this class of mechanisms is simple to reconcile with narrow inference. Under narrow inference, the agent perceives the transfer as measurable only with respect to their own actions. Suppose the principal could choose a more general mechanism in which the transfer function varied with an arbitrary message space as well as the actions. The principal could present information on how the transfer varies with more finely grained messages, drawing the agent's attention to the joint multidimensional nature of their problem and undoing the narrow inference.

In Section 3, I show the restriction makes no difference to the analysis of the principal's optimal mechanism in the rational case. Under the optimal mechanism the agent chooses a strategy that is deterministic, and as such can be implemented with a transfer function that only depends on the chosen action.

2.2 Model Interpretations

The model allows actions to have both a positive and negative effect on payoffs. The sign of $v_i(s)$ determines whether a type s agent has predictable positive utility from action $a_i = 1$ or predictable disutility. Likewise, the direct effect of actions on the principal's payoff can be positive ($w_i \geq 0$) or negative ($w_i \leq 0$).

This framework can capture the stories given in the introduction. Suppose the agent is a buyer of goods and the principal a seller. In this story, $v_i(s) > 0$ represents the predictable payoff benefit to the type s buyer of purchasing the good $a_i = 1$, while $w_i < 0$ represents the cost to the firm of production. The seller then chooses a pricing schedule t.

In another story the agent is a worker and the principal a firm. Here $v_i(s) < 0$ is the cost of obtaining human capital and $w_i > 0$ is the benefit to the firm of the human capital the worker acquires. The firm then chooses a earnings schedule t, which depends on the skills the workers acquire.

2.3 Rational Inference

Given a strategy g, write the expected utility of an agent of type s as

$$U(s) = \sum_{a \in A} g(a|s)u(s, a, t(a)) = \sum_{a \in A} g(a|s)[\sum_{i \in N} v_i(s)a_i + t(a)] \tag{4}$$

Incentive Compatibility (IC) of strategy g under transfer function $t \in \mathbb{R}^A$ requires that g is a best response to t. This means for any $s \in S$, $a \in supp(g(.|s))$ and any $a' \in A$

$$\sum_{i \in N} v_i(s) a_i + t(a) \geq \sum_{i \in N} v_i(s) a'_i + t(a') \tag{5}$$

The agent always has the option of not participating in the mechanism, taking the actions $a = 0$ and obtaining zero transfer. This gives us the following participation constraint: for all $s \in S$

$$U(s) \geq 0 \tag{6}$$

2.4 Narrow Inference

Given a strategy g, an unconditional distribution over actions in A is induced as follows.

$$g(a) = \int_0^1 g(a|s)\, p(s) ds \tag{7}$$

Let the marginal over an action in dimension i be denoted $g_i(a_i) = \sum_{a_{-i} \in A_{-i}} g(a_i, a_{-i})$. I use the terms action distribution and strategy interchangeably throughout the paper.

The agent forms narrow perceptions of the mechanism's transfer function. In particular an agent believes when taking a decision in dimension i that in expectation they will receive $\bar{t}_i(a_i)$ if they take action a_i. When $g_i(a_i) > 0$ we require that this expectation is consistent with the actual conditional expectation of transfers given a_i.

$$\bar{t}_i(a_i) = \sum_{a_{-i} \in A_{-i}} \frac{g(a_i, a_{-i})}{g_i(a_i)} t(a_i, a_{-i}) \tag{8}$$

Denote the vectors of beliefs $\bar{t}(a) = (\bar{t}_i(a_i))_{i=1}^n \in \mathbb{R}^n$ and $\bar{t} = (\bar{t}(a))_{a \in A} \in \mathbb{R}^{2n}$. When $g_i(a_i) = 0$, we allow $\bar{t}_i(a_i)$ to take on arbitrary values. Analogously to the rational benchmark, these 'off-path' actions do not affect the principal's objective and $\bar{t}$ can be set to ensure incentive compatibility. Henceforth, I refer to agents performing narrow inference as 'narrow agents'.

A narrow agent imposes an additively separable form on their estimate of the transfer function using $\bar{t}$. This gives them the following perceived expected utility from strategy g when they are of type $s \in S$.

$$\overline{U}(s) = \sum_{i \in N} \sum_{a_i \in A_i} g_i(a_i|s)[v_i(s)a_i + \overline{t}_i(a_i)] = \sum_{i \in N} \overline{U}_i(s) \tag{9}$$

Where

$$\overline{U}_i(s) = \sum_{a_i \in A_i} g_i(a_i|s)[v_i(s)a_i + \overline{t}_i(a_i)] \tag{10}$$

denotes the narrow perceived expected utility of type s in dimension i. A strategy g is *narrow incentive compatible* (NIC) if for any dimension $i \in N$, type $s \in S$ and actions $a_i \in supp(g_i(.|s))$, $a_i' \in A_i$

$$v_i(s)a_i + \overline{t}_i(a_i) \geq v_i(s)a_i' + \overline{t}_i(a_i') \tag{11}$$

We assume there are dimension by dimension *narrow participation constraints*: for all $s \in S$, $i \in N$

$$\overline{U}_i(s) \geq 0 \tag{12}$$

This fits the following interpretation; when facing a decision on a given dimension, the agent's perceived utility from continuing to participate and taking their best action on that dimension has to exceed the utility they could obtain from leaving the mechanism at that point — which as in the rational benchmark is normalized to zero. This is in line with the agent believing the true transfer function is additive and not taking a coordinated joint decision on participation.

In Section C of the Online Appendix, I consider an alternative joint participation constraint where the sum of the narrow utilities must exceed zero; $\sum_{i \in N} \overline{U}_i(s) \geq 0$. Any mechanism that satisfies the narrow participation constraints also satisfies this joint participation constraint, and we show how the results hold with modification under the alternative participation constraint.

3 Rational Benchmark

With rational agents, we have a screening problem with a single dimension of type but multiple dimensions of action. I restate existing results from Carroll (2017) adapted to our setting.[3]

[3] In particular Proposition 3.1 of Carroll (2017). The model maps on to the model of that paper with a comonotone type distribution, so that the action $a_i = 1$ on any dimension has a

It will be shown that the agent's strategy under the principal's optimal mechanism with rational agents takes a *threshold form* where there is a potentially different threshold $\hat{s}_i \in S$ on each dimension such that $g_i(1|s) = \mathbb{1}\{s \geq \hat{s}_i\}$. Let the vector of thresholds across dimensions be denoted $\hat{s} = (\hat{s}_i)_{i\in N} \in [0,1]^n$. We can characterize the principal's problem in terms of choosing these thresholds. Denote the value of the principal's objective under threshold strategy $\hat{s}$ by $W(\hat{s})$.

Proposition 1. *Assume the IVV assumption holds. The principal maximizes their objective over all IC mechanisms that satisfy the participation constraint if and only if they choose a transfer function implementing a threshold strategy that solves the following problem.*

$$\max_{\hat{s}\in[0,1]^n} W(\hat{s}) = \sum_{i\in N} \int_{\hat{s}_i}^{1} (\phi_i(s) + w_i)\, p(s)\, ds \tag{13}$$

The principal's value under an objective maximizing mechanism can be achieved by an additively separable transfer function

$$t(a_1, ..., a_n) = \sum_{i\in N} t^i(a_i) \tag{14}$$

$$t^i(0) = 0, t^i(1) = -v_i(\hat{s}_i) \text{ for all } i \in N \tag{15}$$

Proof. See Online Appendix Section B □

Thus, the principal's optimal transfer function can be treated as the sum of separate transfer functions, one for each dimension. This does not result directly from IC, but rather from the optimality for the principal of implementing a threshold strategy when IVV holds. We see in Section D of the Online Appendix that without IVV it can be optimal for the principal to choose a non-separable transfer function, and that under such a transfer function a non-threshold strategy is implemented and thus IC.

predictable utility $q \in [v_i(0), v_i(1)]$ that is distributed according to cdf $P(v_i^{-1}(q))$. Earlier work by Mirman and Sibley (1980) analyzes the monopoly screening model with many dimensions of action but a single dimension of type, but assumes directly that the IC action on any dimension is increasing in the type. In Carroll (2017) this is explicitly shown to hold under the regularity (IVV) assumption.

4 Narrow Agents

The solution to the principal's design problem with narrow agents can be characterized as a zero-sum game between the principal and an adversarial player. In this game, the principal faces the same design problem as in the rational benchmark except the predictable utilities are scaled by some factor $\beta_i \in [0,1]$ in each dimension. The predictable utility in dimension i is then $\beta_i v_i(s)$, and the scaling factors sum to one across dimensions $\sum_{i \in N} \beta_i = 1$. The principal chooses a mechanism to maximize their objective while the adversarial player simultaneously chooses the scaling factors to minimize the value of the objective. The result shows that the agent's strategy and value of the principal's objective under the principal's optimal mechanism with narrow agents coincide with those that arise as the solution to this zero sum game with rational agents.

An interpretation of the shrinkage factors is as follows. For a given perceived size of incentives, you have a transfer function that would achieve these perceived incentives under narrow inference, and a transfer function that would achieve these perceived incentives under the rational benchmark. The shrinkage factors give how much these transfers for the rational benchmark would have to be scaled to achieve the same expected transfer as under narrow inference. The fact the shrinkage factors must be between zero and one and sum to one demonstrates that narrow inference amplifies the size of incentives relative to the rational benchmark.

4.1 Main Characterization Result

Theorem 1. *Assume the IVV assumption holds. The principal maximizes their objective over all NIC mechanisms that satisfy the narrow participation constraints if and only if they choose a transfer function that implements a threshold strategy that solves*

$$\min_{\beta \in [0,1]^n : \sum_{i \in N} \beta_i = 1} \; \max_{\hat{s} \in [0,1]^n} \overline{W}(\hat{s}; \beta) = \max_{\hat{s} \in [0,1]^n} \; \min_{\beta \in [0,1]^n : \sum_{i \in N} \beta_i = 1} \overline{W}(\hat{s}; \beta) \tag{16}$$

with the value of the principal's objective given by

$$\overline{W}(\hat{s}; \beta) = \sum_{i \in N} \int_{\hat{s}_i}^{1} (\beta_i \phi_i(s) + w_i)\, p(s) ds \tag{17}$$

Proof. In Appendix □

To see some intuition for this result, consider the case with two dimensions and symmetric predictable utility $v_i(s) = v(s)$ and principal's direct utility $w_i = w$ from actions across all dimensions $i \in N$. With a rational agent, from Proposition 1 the principal's optimum sets a threshold $\hat{s}$ so that the action is taken on all dimensions for all types above and the zero action is taken on all dimensions for all types below. This optimum is induced with a transfer function that is additive and identical across dimensions $t(a_1, a_2) = t^1(a_1) + t^2(a_2)$ and $t^1(1) = t^2(1) = \tilde{t}(1)$, $t^1(0) = t^2(0) = 0$. With a narrow agent, under the same strategy and transfer function the agents double count the effect of each action on the transfer. In each dimension, they believe that the transfer resulting from taking the action $a_i = 1$ is $2 \cdot \tilde{t}(1)$ and the transfer resulting from $a_i = 0$ is 0. This double-counting is the result of confounding neglect; the agent fails to adjust for the fact that every type who takes action $a_1 = 1$ also takes action $a_2 = 1$. The principal then has to halve the size of the difference in transfers in order to maintain the same thresholds $\frac{1}{2}\tilde{t}(1)$. This has the same effect as scaling the predictable utilities down by $\frac{1}{2}$ in each dimension when the agent is rational.

The result extends this logic to asymmetric cases. It allows us to both solve for the principal's optimal mechanism with narrow agents and also demonstrates the connection between any given problem with narrow agents to the rational benchmark. I use the characterization to obtain additional results. I give conditions under which the principal does and does not benefit from facing narrow over rational agents, and how the implemented strategy changes between the two cases. I then explore the effect of merging and splitting the costs and benefits of actions across dimensions and what happens when the number of dimensions grows large. I now present some preliminaries that are used in obtaining the characterization.

4.2 Preliminaries for Characterization Result

I first consider which beliefs can be induced by some transfer function. I show that beliefs must satisfy a statistical correctness constraint, and that for any threshold strategy there is a valid additively separable transfer function that induces beliefs satisfying this statistical correctness constraint.

It will be useful in characterizing the principal's optimal mechanism under NIC to show how beliefs and the transfer function relate given any fixed distribution over actions. The following result shows when we can write the transfer distribution in terms of the beliefs over the expected transfer in either dimension. It gives a standard statistical correctness result that applies to beliefs formed using Bayesian Networks under perfect Direct Acyclic Graphs (DAGs) (Spiegler, 2020).

Lemma 1 (Statistical Correctness)**.** *Given any distribution over actions g, for any two dimensions $i, j \in N$ we have that beliefs $\bar{t}_i, \bar{t}_j$ satisfy*

$$\sum_{a_i \in A_i} g_i(a_i)\bar{t}_i(a_i) = \sum_{a_j \in A_j} g_j(a_j)\bar{t}_j(a_j) \tag{18}$$

Proof. Rearranging the expected transfer gives us the first part. For any $i \in N$:

$$\sum_{a \in A} g(a)t(a) = \sum_{a_i \in A_i} g_i(a_i) \sum_{a_{-i} \in A_{-i}} \frac{g(a_i, a_{-i})}{g_i(a_i)} t(a_i, a_{-i}) = \sum_{a_i \in A_i} g_i(a_i)\bar{t}_i(a_i)$$

□

For a fixed strategy and action distribution, this statistical correctness is necessary but not sufficient for a transfer function to exist that induces given beliefs. For an example of beliefs that satisfy the statistical correctness constraint but cannot be induced, consider the case with $N = \{1, 2\}$ and $g(1, 1) = g(0, 0) = \frac{1}{2}$. If beliefs do not also satisfy $\bar{t}_1(1) = \bar{t}_2(1)$ and $\bar{t}_1(0) = \bar{t}_2(0)$, then there is no transfer function implementing these beliefs under this action distribution.

The principal's objective can be written in terms of beliefs. This means it is useful to work directly with beliefs rather than the underlying transfer function when we characterize the principal's optimal mechanism. Although the statistical correctness constraint is not sufficient, it will be sufficient if the strategy of the agent takes a threshold form. I now show that any NIC strategy must take a threshold form. This differs from the rational case where a threshold strategy is optimal for the principal under the IVV assumption, but is not an implication of IC.

Lemma 2. *Every NIC strategy takes a* ***threshold form*** *for almost all $s \in S$.*

Any strategy g that takes a ***threshold form*** *is NIC if there exists a transfer function t that together with g induces beliefs that for all $i \in N$ satisfy*

$$\bar{t}_i(1) = \bar{t}_i(0) - v_i(\hat{s}_i) \tag{19}$$

Proof. Let $\hat{U}_i(s, a_i) = v_i(s)a_i + \bar{t}_i(a_i)$. Clearly $\hat{U}_i(s, 1) - \hat{U}_i(s, 0) = v_i(s) + \bar{t}_i(1) - \bar{t}_i(0)$ is increasing in s. There is a threshold $\hat{s}_i$ such that $\hat{U}_i(s, 1) \geq \hat{U}_i(s, 0)$ if and only if $s \geq \hat{s}_i$, in which case type s will choose $a_i = 1$. Given a threshold strategy, the beliefs in the proposition statement ensure that the agent is indifferent between taking either action at the threshold. □

The next result shows how any threshold strategy can be made NIC by a transfer function that is additive across dimensions. This means a transfer function exists that implements any beliefs that satisfy statistical correctness, if the agent is choosing actions according to a threshold strategy.

Proposition 2. *For any threshold strategy g, we can construct a transfer function t that induces beliefs $\bar{t}$ so that g is NIC. The constructed transfer function is additive; for any $a_{-i}, \tilde{a}_{-i} \in A_{-i}$ we have that*

$$t(1, \tilde{a}_{-i}) - t(0, \tilde{a}_{-i}) = t(1, a_{-i}) - t(0, a_{-i}) \tag{20}$$

Moreover, any transfer function $\tilde{t}$ such that g is NIC can only differ from this additive t at action combinations that do not occur under g; $t(a) \neq \tilde{t}(a)$ only if $g(a|s) = 0$ for all $s \in S$.

Proof. In Appendix □

The principal can exploit two features of the narrow agent's misperception; that they can only perceive of the transfer function as additive and that their beliefs do not account for confounding bias. Proposition 2 means that if the principal implements a threshold strategy, they have no payoff gain from implementing a transfer function that is not additive. Thus, the principal does not exploit the agent's potentially false perception of the transfer function as additive in their choice of optimal mechanism.

The transfer function can be constructed as additive because under a threshold strategy, there are action combinations that are not chosen by any type of agent. Take the case with two dimensions of action, and assume that the agent's strategy is such that threshold for the first dimension is lower than the threshold in the second dimension; $\hat{s}_1 < \hat{s}_2$. Types in interval $[0, \hat{s}_1)$ choose actions $(0, 0)$, types in $[\hat{s}_1, \hat{s}_2)$ choose $(1, 0)$ and types in $[\hat{s}_2, 1]$ choose $(1, 1)$. The transfer for actions $(0, 1)$, chosen by no type, can be set so that the transfer function is additive. The case where $\hat{s}_1 \geq \hat{s}_2$ is symmetric. Proposition 2 extends this logic to the general case with any number of dimensions. The transfer function is constructed recursively so as to induce the given beliefs.

The proof of Theorem 1 works as follows. First, using Proposition 2, we write both the principal's objective and the statistical correctness constraint from Lemma 1 only in terms of the threshold strategy and the narrow perceived transfer of the type taking action zero on one of the dimensions; $\bar{t}_i(0)$. We then show we can obtain an upper bound to the principal's problem that takes a max-min form, with the β weights in the proof coming from a rewriting of the Lagrange multipliers from the statistical correctness constraint. Under IVV[4], we can apply a standard minimax theorem argument to obtain a saddle point, allowing us to interchange the min and the max. Finally, we show that this minimax upper bound can be attained in the principal's full problem.

4.3 Effect of Narrow Agents on the Principal's Welfare

Using the characterization of the principal's optimal mechanism, I show that when actions have predictable costs for all types of the agent then the principal is better off under narrow inference. Conversely, when actions have predictable benefits for all types then the principal is worse off.

Proposition 3. *Assume the IVV assumption holds.*

1. *When for every $i \in N$ we have $v_i(s) \leq 0$ for all $s \in S$, the principal can obtain at least as high an objective value when the agent is narrow compared to the rational benchmark.*

2. *When for every $i \in N$ we have $v_i(s) \geq 0$ for all $s \in S$, the principal obtains at least as high an objective value in the rational benchmark compared to when the agent is narrow.*

Proof. In Appendix □

The intuition for this result is as follows. When the agent's action is costly in terms of the agent's predictable utility, for any threshold strategy the principal makes transfers to the agent. The single dimension of type results in actions in one dimension being positively correlated with actions on any other dimension. Since actions result in higher transfers to the agent, this leads a narrow agent to overestimate the transfer they will get from the principal.

[4] Without IVV, we can obtain a modified version of the result by using the ironing technique of Myerson (1981). Details of how this work can be found in Online Appendix Section D

Rational agents adjust for the fact their type is on the margin between taking an action or not, so know they will receive a lower transfer than the average obtained by agents taking that action. The overestimation of the transfer by narrow agents means less transfer has to be given to higher type agents in order to implement any given strategy. When actions have predictable utility benefits to the agent, the principal is a net receiver of transfers from the agent. For clearer intuition, consider the application where the principal is a profit-maximizing seller of multiple goods and the agent is a customer. Narrow inference results in the agent overestimating the price of the good, which hurts the seller as it means they sell to fewer types at any price schedule.

We can then see how the principal's optimal thresholds differ when we move to narrow agents from the rational benchmark. Under narrow inference, if actions have a predictable cost to the agent but benefit to the principal, the principal implements a strategy with a lower type threshold for taking the action on any dimension than in the rational benchmark. The opposite holds in the predictable utility benefit, principal's loss case where the thresholds are higher when the agent is narrow.

Proposition 4. *Assume the IVV assumption holds.*

1. *When for every $i \in N$ we have $w_i > 0$ and $v_i(s) \leq 0$ for all $s \in S$, then on each dimension the objective-maximizing thresholds are weakly lower with narrow agents than under the rational benchmark, so $a_i = 1$ is taken by a larger proportion of types for all $i \in N$.*

2. *When for every $i \in N$ we have $w_i < 0$ and $v_i(s) \geq 0$ for all $s \in S$, then on each dimension the objective-maximizing thresholds are weakly greater with narrow agents than under the rational benchmark, so $a_i = 1$ is taken by a smaller proportion of types for all $i \in N$.*

Proof. In Appendix □

Following the same intuition as for Proposition 3, in the first case narrow inference reduces the marginal cost to the principal of implementing that any given proportion of agents take the action on any dimension. Given the fixed benefits of the actions to the principal, this lower marginal cost means the principal wants a higher proportion of agents to take the action. In the second case, where actions predictably cost the principal but benefit the agent the principal has a lower marginal benefit from a higher proportion of agents taking the action, but a fixed cost. The principal then wants to reduce the proportion of agents taking the actions on all dimensions.

Suppose that whether the agent does narrow inference is something that the principal can influence or design. This could be either through the way they present the mechanism or data about the mechanism. If the default is that the agent does narrow inference, the principal can try to educate the agent on how to do rational inference. Proposition 3 then shows in what cases the principal would gain from de-biasing the agent and in what cases they would not. In settings like the firm-worker story, where the agent makes costly investments that benefit the principal, then the principal wants to salami slice the agent's investment choices into smaller decision problems as much as possible. Conversely, in settings like the buyer-seller story where the agent buys valuable goods that are costly for the principal to produce, the principal wants to merge choices into few decisions. This presents a novel motivation for the seller to engage in bundling the goods they are selling, that aims to mitigate the costs of the buyers agent's bounded rationality to the seller rather using monopoly power to exploit the buyer.[5] I explore the idea of merging and splitting dimensions more formally in the next section.

4.4 Splitting and Merging Dimensions

The ability to add or subtract new dimensions, and thus to combine or split the costs of existing actions, can have an impact under narrow inference through the agent's beliefs. We say that a dimension space M is a *split* of $N \subset M$ if there is a partition of M according to the dimensions in N where for each $i \in N$, there is a partition cell $\mathcal{P}(i)$ so that $v_i(s) = \sum_{j \in \mathcal{P}(i)} v_j(s)$ and $w_i = \sum_{j \in \mathcal{P}(i)} w_j$. We also refer to N as a *merge* of dimensions in M, and denote the cardinality of N by $n = |N|$ and the cardinality of M by $m = |M|$.

Under the rational benchmark, splitting dimensions can only benefit the principal as it enlarges the set of strategies that can be implemented. In contrast, under narrow inference splitting and merging dimensions affects the agent's perceptions of the transfers. This means the impact of these dimensional changes depends on whether actions have predictable benefits or costs to the agent and the principal.

[5] See Armstrong (2016) for a review of the literature on nonlinear pricing and bundling.

Proposition 5. *Assume the IVV assumption holds and the agent does narrow inference.*

1. *Given any split M of dimension space $N \subset M$, if for every $i \in N$ and $j \in M$ we have $w_i > 0$, $w_j > 0$ and $v_i(s) \leq 0$, $v_j(s) \leq 0$ for all $s \in S$, the principal can obtain at least as high an objective value under M as under N.*

2. *Given any split M of dimension space $N \subset M$, if for every $i \in N$ and $j \in M$ we have $w_i < 0$, $w_j < 0$ and $v_i(s) \geq 0$, $v_j(s) \geq 0$ for all $s \in S$, the principal can obtain at least as high an objective value under N as under M.*

Proof. In Appendix □

4.4.1 Effect of Large Dimensionality

I now consider what happens when the costs and benefits of the actions are split across a large dimension space, with symmetry across dimensions. This symmetry means that there is no advantage or cost to splitting dimensions in the rational benchmark. Thus in this setting we can isolate the effects of splitting and merging dimensions that arises from distorted perceptions under narrow inference.

Define a *symmetric dimension space of size n* as follows. For any n, $v_i^{(n)}(s) = \frac{1}{n}v(s)$, and $w_i = \frac{1}{n}w$. Both predictable utility and the principal's direct utility from actions are decreasing as the dimensionality of the action space n grows, but such that the total effect of actions $\sum_{i \in N} v_i^{(n)}(s) = v(s)$, $\sum_{i \in N} w_i^{(n)} = w$ is constant. Let $\hat{s}_i^{(n)}$ be the solution to the principal's problem with narrow agents when there is a symmetric dimension space of size n.

We can show that for a symmetric dimension space, under both the rational benchmark and under narrow inference the principal wants to implement that thresholds are equalized.

Lemma 3. *Assume the IVV assumption holds. Then for any symmetric dimension space of size n, we have that in both the rational benchmark and under narrow inference the principal wants to equalize the implemented thresholds across dimensions: $\hat{s}_i = \hat{s}_j$ for any $i, j \in N$.*

Proof. In Appendix □

Now we consider what happens as n grows large. In this setting when actions have a direct benefit for the principal, the principal is able to incentivize types to take the actions at vanishing cost. This is because they only have to pay the transfer to the agent on one dimension in this symmetric narrow agent setting. When actions have a direct cost to the principal, the opposite is true and it becomes too costly for the principal to extract transfers from the agent. In this case, narrow agents overestimate the transfer cost to themselves of taking the action for any given transfer function.

Proposition 6. *Assume the IVV assumption holds. Consider a sequence as $n \to \infty$ of symmetric dimension spaces of size n.*

1. *When $w > 0$, there exists an $\overline{n}$ such that for any $n \geq \overline{n}$, we have that the principal's optimal mechanism with narrow agents implements a strategy such that all types take the action on all dimensions: for all $i \in N$, $\hat{s}_i^n = 0$.*

2. *When $w < 0$, there exists an $\overline{n}$ such that for any $n \geq \overline{n}$, we have that the principal's optimal mechanism with narrow agents implements a strategy such that all types take no action on all dimensions: for all $i \in N$, $\hat{s}_i^n = 1$.*

Moreover as $n \to \infty$, the expected transfer between the agent and the principal shrinks to zero: $\sum_{a \in A} g(a)t(a) \to 0$.

Proof. In Appendix □

The result follows from the logic discussed in the intuition for the characterization result in Section 4.1. To implement symmetric thresholds, the principal must scale down the transfer utility from taking the action on any dimension by $\frac{1}{n}$ relative to the rational benchmark to maintain the same thresholds as narrow incentive compatible. As n grows large, the contribution of the transfer to the principal's utility then shrinks and the transfers to or from the agent become smaller. Eventually for some n, the direct predictable utility to the principal dominates their objective. When $w > 0$ this means they want all types of agent to take the beneficial action while when $w < 0$ they want no types to take the action.

4.5 The Shape of the Principal's Optimal Mechanism

We have seen in the symmetric case that when actions directly benefit the principal, then as the number of dimensions grows towards infinity the principal converges to being able to extract the full surplus of the interaction with the agent.

The reason for this is that the principal is able to load the transfer cost of incentivizing the actions onto only one of the dimensions. They can then rely on the agent over-counting the transfer benefits on the other dimensions to ensure that all types of agent undertake the costly actions.

I now show that the transfer function does not only take this form in this special case. An implication of Theorem 1 is that the principal's optimum can be implemented with a transfer function that takes this one-dimensional loaded form in general.

Proposition 7. *Assume the IVV assumption holds. Take the welfare achieved by the Principal under Theorem 1. There must be a solution that attains this value; $(\hat{s}^*, \beta^*) \in [0,1]^n \times [0,1]^n$, such that the threshold strategy $\hat{s}^*$ is NIC and satisfies the narrow participation constraints under the following additive transfer function, for some $i^* \in N$*

$$t(a_1, ..., a_n) = \sum_{i \in N} t^i(a_i) \tag{21}$$

$$t^{i^*}(1) = -v_{i^*}(\hat{s}^*_{i^*}) \tag{22}$$

$$t^j(1) = t^j(0) = 0 \text{ for all } j \in N \setminus \{i^*\} \tag{23}$$

Proof. In Appendix □

Note that not all NIC strategies and beliefs satisfying the narrow participation constraints can be implemented with a transfer function that has these properties. Consider $N = \{1, 2\}$ and implementing the strategy $0 < \hat{s}_1 < \hat{s}_2 < 1$ with beliefs $(\bar{t}_1(0), \bar{t}_2(0)) = (0, v_2(\hat{s}_2)(1 - P(\hat{s}_2)) - v_1(\hat{s}_1)(1 - P(\hat{s}_1)))$, $\bar{t}_j(1) = \bar{t}_j(0) - v_j(\hat{s}_j)$ where $v_2(\hat{s}_2)(1 - P(\hat{s}_2)) - v_1(\hat{s}_1)(1 - P(\hat{s}_1)) > 0$. From Proposition 2, an additive transfer function that can do so must satisfy

$$t(a_1, a_2) = t^1(a_1) + t^2(a_2) \tag{24}$$

$$t^1(1) - t^1(0) = \frac{P(\hat{s}_2)}{P(\hat{s}_2) - P(\hat{s}_1)}(v_2(\hat{s}_2)(1 - P(\hat{s}_2)) - v_1(\hat{s}_1)(1 - P(\hat{s}_1))) \tag{25}$$

$$t^2(1) - t^2(0) = -v_2(\hat{s}_2) - \frac{P(\hat{s}_1)}{P(\hat{s}_2) - P(\hat{s}_1)}(v_2(\hat{s}_2)(1 - P(\hat{s}_2)) - v_1(\hat{s}_1)(1 - P(\hat{s}_1))) \tag{26}$$

This transfer function clearly cannot in general take the one-dimensional loaded form of Proposition 7.

4.6 Illustrative Example

The following example illustrates the rational benchmark and the narrow agent results.

Example 2. We return to the firm-worker story. A worker faces two human capital investment decisions $N = \{1, 2\}$. They can choose to obtain technical skills $a_1 = 1$ or not $a_1 = 0$ and/or to obtain managerial skills $a_2 = 1$ or not $a_2 = 0$. The disutility of the worker from obtaining the dimension i skills is given by $v_i(s) = -r_i(1 - s)$. This varies with a uniformly distributed type; $s \sim U[0, 1]$, $P(s) = s$. These skills have equal benefit for the firm that employs the worker; $w_1 = w_2 = w > 0$. Assume that it is more costly for the agent to acquire management skills than to acquire technical skills, $r_2 \geq r_1 > 0$.

For these parameters, we have that for each $i \in N$

$$\phi_i(s) = v_i(s) - \frac{1 - P(s)}{p(s)} v_i'(s) = -2r_i(1 - s) \tag{27}$$

We have increasing virtual values and the strategy implemented by the principal's optimal mechanism has a threshold form $g_i(1|s) = \mathbb{1}\{s \geq \hat{s}_i^*\}$ in each dimension, with thresholds

$$\hat{s}_i^* = (1 - \frac{w}{2r_i})\mathbb{1}\{w \leq 2r_i\} \tag{28}$$

This illustrates some simple corollaries of the principal's optimal mechanism in the rational benchmark given by Proposition 1. The threshold in any dimension only depends on the predictable utilities for that dimension, and is decreasing in the size of the predictable benefit to the principal w and increasing in the predictable cost to the agent r_i. These thresholds give the principal objective value

$$W^{rat} = \sum_{i \in N} [\frac{w^2}{4r_i}\mathbb{1}\{w \leq 2r_i\} + (w - r_i)\mathbb{1}\{w > 2r_i\}] \tag{29}$$

With an agent who does narrow inference, given β, the thresholds that solve the problem $\max_{\hat{s} \in [0,1]^n} \overline{W}(\hat{s}, \beta)$ are

$$\hat{s}_i(\beta_i) = (1 - \frac{w}{2r_i\beta_i})\mathbb{1}\{w \leq 2\beta_i r_i\} \tag{30}$$

for $i \in N$. These are always weakly less than the rational thresholds (28), which demonstrates Proposition 4.

From Theorem 1, we can solve for the principal's optimal thresholds and the value of the principal's objective. There are four separate cases

1. If $w \geq 2r_2$ then $\hat{s}_1 = \hat{s}_2 = 0$ is optimal for the principal, and the principal obtains welfare
$$W_1 = -r_2 + 2w$$

2. If $2r_2 > w \geq 2\sqrt{r_1 \cdot r_2}$ then $\hat{s}_1 = 0, \hat{s}_2 = 1 - \frac{w}{2r_2}$ is optimal for the principal, and the principal obtains welfare
$$W_2 = \frac{w^2}{4r_2} + w$$

3. If $2\sqrt{r_1 r_2} > w \geq \frac{2\sqrt{r_1 \cdot r_2}}{1+\sqrt{\frac{r_2}{r_1}}}$ then $\hat{s}_1 = 0$, $\hat{s}_2 = 1 - \sqrt{\frac{r_1}{r_2}}$ is optimal for the principal, and the principal obtains welfare
$$W_3 = -r_1 + \sqrt{\frac{r_1}{r_2}}w + w$$

4. If $\frac{2\sqrt{r_1 \cdot r_2}}{1+\sqrt{\frac{r_2}{r_1}}} > w \geq 0$ then $\hat{s}_1 = 1 - \frac{w}{2r_1}(1 + \sqrt{\frac{r_1}{r_2}})$, $\hat{s}_2 = 1 - \frac{w}{2r_2}(1 + \sqrt{\frac{r_2}{r_1}})$ is optimal for the principal, and the principal obtains welfare
$$W_4 = \frac{w^2}{4}(\frac{1}{\sqrt{r_1}} + \frac{1}{\sqrt{r_2}})^2$$

We can compare the principal's objective value under the rational benchmark W^{rat} with the objective value for narrow agents in any of the four regimes. We can verify that for the parameter values under which each case is optimal, the principal's welfare is higher under narrow inference than in the rational benchmark. This demonstrates Proposition 3.

Note also that the agent's predictable utility in one dimension can affect the optimal threshold in the other dimension. This is particularly clear when looking at the third case, where a small increase in the cost of acquiring technical skills r_1 can actually lead to an increase in the proportion of types the principal wants to obtain managerial skills. This is in contrast to in the rational benchmark, where under IVV the predictable utility of both the agent and the principal only affect the threshold in their own dimension and where less predictable utility for any action means a higher threshold is optimal for the principal. △

5 Discussion

5.1 Testable Implications of Design for Narrow Inference

As demonstrated in Section 4.5, under narrow inference the principal can implement their optimum by choosing a transfer function that only depends on actions chosen in one dimension. This provides a justification for firm hierarchies that only reward one particular observed dimension of worker performance or pricing policies that only charge for one particular product dimension for a complex multi-dimensional item. Empirically, we often observe these simple incentive schemes in settings where more complicated structures could be designed.

The results on splitting and merging dimensions in Section 4.4 imply that organizations may want to present their workers with complicated multi-dimensional training options or present promotion opportunities in terms of many different disaggregated qualifications rather than a few simple ones that may embody the same skill content. In contrast they suggest a firm selling a product wants to make the purchasing choices of a consumer as simple and low-dimensional as possible. Similarly, a firm faced with a worker who makes narrow inference would want to present a non-transparent wage structure where the worker can only obtain partial data on the wage benefits of taking different roles in the firm. A selling firm would want to try and induce rational behaviour by the consumer by providing transparent data on how prices jointly vary with product dimensions.

5.2 Future Research Directions

One can imagine many additional variations and extensions of the model of bounded rationality explored in this paper. In particular, it seems interesting to consider how the principal could shape the extent of agent's departure from rational beliefs. For example, the principal could provide data on how additional variables correlate with the transfers or otherwise frame the mechanism in a way that influences inference by the agent. The analysis of this paper suggests that in some cases this could be as important a margin of design as the size of material incentives.

Another direction would be to consider the model of narrow inference outside of the principal-agent context studied in this paper. For example, labour-market settings where many firms compete to employ workers, or industrial organization settings where many firms sell products. It would be interesting to study how narrow inference by workers or consumers shapes wage and price setting in these competitive settings.

A Appendix

A.1 Proofs

Proof of Proposition 2

By Lemma 2, the beliefs inducing any threshold strategy must satisfy

$$\bar{t}_i(1) = \bar{t}_i(0) - v_i(\hat{s}_i)$$

for each dimension $i \in N$. Thus NIC and the thresholds pin down beliefs, and any threshold strategy can be rendered NIC by some beliefs.

An outline of the proof is as follows. In Lemma A.1 we show that for any two distinct action combinations that occur under a threshold strategy, one of the action vectors is weakly larger on all dimensions. For each dimension we can partition the set of action combinations, with each cell consisting of all action combinations that share a common action for that dimension. In Lemma A.2 we then show that for every action combination that occurs with positive probability, there is at least one dimension such that every other action combination in the same partition cell for that dimension has a smaller action taken on at least one of the other dimensions.

We can use this fact to recursively construct a transfer function that implements given beliefs, and we can show that this constructed transfer function is additive for all action combinations that occur with positive probability. Finally we extend this transfer function so that is defined on all action combinations, whilst preserving additivity.

Lemma A.1. *Let $\tilde{g}$ be a threshold strategy. Then for any $s > s'$ and a'', a' such that $a'' \neq a'$, $\tilde{g}(a''|s'') > 0$ and $\tilde{g}(a'|s') > 0$ only if $a''_j \geq a'_j$ for all $j \in N$.*

Proof. Since $\tilde{g}$ is a threshold distribution, for any two dimensions $i, j \in N$ there is an $\hat{s}_i$ such that $\tilde{g}_i(1|s) = \mathbb{1}\{s \geq \hat{s}_i\}$ and an $\hat{s}_j$ such that $\tilde{g}_j(1|s) = \mathbb{1}\{s \geq \hat{s}_j\}$. Suppose that we can find $a'' \neq a'$, $\tilde{g}(a''|s'') > 0$ and $\tilde{g}(a'|s') > 0$ for some $s'', s' \in S$ such that $a''_i = 1 > a'_i = 0$ for some dimension $i \in N$ but $a''_j = 0 < a'_j = 1$ for another dimension $j \in N \setminus \{i\}$. But then $s \geq \hat{s}_i > s''$ and $s'' \geq \hat{s}_j > s$, a contradiction. □

This implies that any two action combinations occurring with positive probability under a threshold strategy can be ranked. Denote the set of all action combinations that have positive probability under g by $A(g) = \{a \in A : \exists s \in S$ such that $g(a|s) > 0\}$. Denote the projection of $A(g)$ on dimension $i \in N$ by $A_i(g)$. Define the order $\succ$ so that $a'' \succ_i a'$ if and only if $a''_j \geq a'_j$ for all $j \in N$ with strict inequality for at least one such j. By Lemma A.1 this is a strict total order.

Given our NIC threshold strategy g, we enumerate the set $A(g) = \{1, ..., |A(g)|\}$ so that $k > l$ means that for $a^k, a^l \in A(g)$, $a^k \succ a^l$. Now we can form a partition of $A(g)$ for each dimension $i \in N$. For each action $a_i \in A_i$, define the set $\mathcal{A}_i(a_i) = \{(a_i, \tilde{a}_{-i}) \in A(g)\}$. This is a partition as $\varnothing \notin \mathcal{A}_i(a_i)$ for any $a_i \in A_i(g)$, $\cup_{\tilde{a}_i \in A_i} \mathcal{A}_i(\tilde{a}_i) = A(g)$ and $\mathcal{A}_i(a''_i) \cap \mathcal{A}_i(a'_i) = \varnothing$ for any $a''_i \neq a'_i \in A_i$.

We can then show that any action combination that occurs with positive probability under an NIC threshold strategy must be maximal in the partition cell according to the order $\succ$ for at least one dimension.

Lemma A.2. *Given an NIC threshold strategy g, any $a \in A(g)$ is such that for at least one dimension $j \in N$, $a = (a_j, a_{-j}) \succ \tilde{a} = (a_j, \tilde{a}_{-j})$ for all $\tilde{a} \in \mathcal{A}_j(a_j)$.*

Proof. Suppose this does not hold, then for all $i \in N$, we can find an $\tilde{a}(i) \in \mathcal{A}(a_i)$ such that $\tilde{a}(i) \succ a$. The finite set $\{\tilde{a}(1), ..., \tilde{a}(n)\}$ must contain a member that is minimal in the strict total order $\succ$. Denote this element $\tilde{a}(k)$ for some $k \in N$. Then on all dimensions $j \in N$, $\tilde{a}(k)_j \leq a_j = \tilde{a}(j)_j$, as if $\tilde{a}(k)_j > a_j = \tilde{a}(j)_j$ then $\tilde{a}(j) \not\succ \tilde{a}(k)$ which would contradict the minimality of $\tilde{a}(k)$ in $\{\tilde{a}(1), ..., \tilde{a}(n)\}$. However, $\tilde{a}(k)_j \leq a_j$ for all $j \in N$ contradicts that $\tilde{a}(k) \succ a$. □

Now using this, for any action combination $a^l \in A(g) = \{1, ..., |A(g)|\}$ assign a dimension $\pi(l) \in N$ so that $a^l \succ \tilde{a}$ for any $\tilde{a} \in \mathcal{A}_{\pi(l)}(g)$. Then we can recursively define a transfer function t from the beliefs $\bar{t}$ that render g NIC. For any $k \in \{2, ..., |A(g)|\}$

$$t(a^1) = \bar{t}_{\pi(1)}(a^1_{\pi(1)})$$

$$t(a^k) = \frac{\sum_{a \in \mathcal{A}_{\pi(k)}(a^k_{\pi(k)})} g(a)}{g(a^k)} \bar{t}_{\pi(k)}(a^k_{\pi(k)}) - \sum_{a \in \mathcal{A}_{\pi(k)}(a^k_{\pi(k)}) \setminus \{a^k\}} \frac{g(a)}{g(a^k)} t(a)$$

This transfer function is well defined as for any $a^k \in A(g)$ all $a \in A_{\pi(k)}(a^k_{\pi(k)})$, $t(a)$ has been defined at an earlier stage as a^k is maximal in $\succ$ on dimension $\pi(k)$.

Since a^1 is minimal in $A(g)$, we have that $a^1 = A_{\pi(1)}(a^1_{\pi(1)})$, so the first equation is in fact a special case of the second. We now show that t is additive for the action combinations $a \in A(g)$ at which it is well defined.

Lemma A.3. *For any $a_{-i} \in A_{-i}$ with $(1, a_{-i}), (0, a_{-i}) \in A(g)$, there exists no $\tilde{a}_{-i} \neq a_{-i}$ such that $(1, \tilde{a}_{-i}) \in A(g)$ and $(0, \tilde{a}_{-i}) \in A(g)$.*

Proof. Suppose for contradiction that there is a $\tilde{a}_{-i} \neq a_{-i}$ such that $(1, \tilde{a}_{-i}) \in A(g)$ and $(0, \tilde{a}_{-i}) \in A(g)$. As we have strict total order $\succ$ on $A(g)$, we have two cases. In the first case $(0, a_{-i}) \succ (0, \tilde{a}_{-i})$. This means that $a_j \geq \tilde{a}_j$ for all $j \in N\backslash\{i\}$ with strict inequality for some such j. Then neither $(0, a_{-i}) \succ (1, \tilde{a}_{-i})$ nor $(1, \tilde{a}_{-i}) \succ (0, a_{-i})$. This is a contradiction since $(1, \tilde{a}_{-i}) \neq (0, a_{-i})$, Lemma A.1 implies they must be ranked.

Similarly if $(0, \tilde{a}_{-i}) \succ (0, a_{-i})$, then $\tilde{a}_j \geq a_j$ for all $j \in N \setminus \{i\}$ with strict inequality for some such j. Then neither $(1, a_{-i}) \succ (0, \tilde{a}_{-i})$ or $(0, \tilde{a}_{-i}) \succ (1, a_{-i})$. □

Therefore we cannot have $t(1, a_{-i}) - t(0, a_{-i}) \neq t(1, \tilde{a}_{-i}) - t(0, \tilde{a}_{-i})$ and $(1, a_{-i}), (0, a_{-i}), (1, \tilde{a}_{-i}), (0, \tilde{a}_{-i}) \in A(g)$. This means the transfer function is additive for all $a \in A(g)$.

We can additively extend the transfer function t defined above to all $a \in A$. Denote this extended transfer function by t'. For any $\tilde{a} \in A$, such that $a^1_j \geq \tilde{a}_j$ for all $j \in N$, define $t'(\tilde{a}) = t(a^1)$. For each dimension $i \in N$, we will define $t^i(a_i)$ for each $a_i \in A_i$ so that $t'(a) = \sum_{i \in N} t^i(a_i)$. First set $t^i(a^1_i) = \frac{1}{n} t(a^1)$ for all $i \in N$, and $t^i(1) = t^i(0)$ if $a^1_i = 1$. For dimensions which are such that $(1, a_{-i}) \notin A(g)$ for every $a_{-i} \in A_{-i}$, set $t^i(1) = t^i(0)$.

Now move through the elements $a^l \in \{2, ..., |A(g)|\} \subset A(g)$ in the $\succ$ order. If a^l differs in one dimension j from a^{l-1}, then by Lemma A.3 there is a unique $a_{-j} \in A_{-j}$ such that $(1, a_{-j}), (0, a_{-j}) \in A(g)$, and we can write $t^j(1) - t^j(0) = t(1, a_{-j}) - t(0, a_{-j}) = t(a^l) - t(a^{l-1})$. If a^l differs from a^{l-1} on multiple dimensions (denoted by the set N^l), choose an arbitrary $j \in N^l$ and set $t^j(1) - t^j(0) = t(a^l) - t(a^{l-1})$ and set $t^k(1) = t^k(0)$ for all other $k \in N^l \setminus \{j\}$. This process results in a transfer function that is additive and such that $t'(a) = \sum_{i \in N} t^i(a_i) = t(a)$ for every $a \in A(g)$.

For the final part, any $a \in A \setminus A(g)$ is such that $g(a) = 0$ and thus $t(a)$ does not affect on-path beliefs $\bar{t}(a_i) = \sum_{a_{-i} \in A_{-i}} \frac{g(a_i, a_{-i})}{g_i(a_i)} t(a)$. For all $a \in A(g)$ and t' that implements the beliefs $\bar{t}$ must match our recursive construction t. To see this take $a^1 \in A(g)$, at this action combination $t'(a^1) = t(a^1)$ is pinned down by the beliefs only. This is because a^1 is the minimal action in $A(g)$ according to $\succ$ but is also maximal in $\succ$ on one of the dimension partitions, so on this dimension $t'(a^1)$ and $t(a^1)$ must both be equal to the belief. Then any expression that implements $\overline{t}$ for $t'(a^2)$ is also pinned down in terms of beliefs according to the recursive formula given for $t(a_2)$. Continuing up the order $\succ$, we have $t'(a^l) = t(a^l)$ for every $a^l \in A(g)$.

Proof of Theorem 1

We break the proof into 3 steps.

Step 1: We write the principal's problem in a virtual value form. From the statistical correctness constraint in Lemma 1 and the result that any NIC strategy takes a threshold form in Lemma 2, for any dimension $i \in N$ we can write

$$\begin{aligned}\int_0^1 \sum_{a \in A} g(a|s)t(a) = \sum_{a_i \in A_i} g_i(a_i)\overline{t}_i(a_i) &= \int_0^1 \sum_{a_i \in A_i} \overline{t}_i(a_i) g_i(a_i|s)\, p(s)ds \\ &= (1 - P(\hat{s}_i))\overline{t}_i(1) + P(\hat{s}_i)\overline{t}_i(0) \\ &= \overline{t}_i(0) - (1 - P(\hat{s}_i))v_i(\hat{s}_i)\end{aligned}$$

For any NIC strategy g inducing thresholds $\hat{s}$. We then have for any $i, j \in N$ that

$$\begin{aligned}&(1 - P(\hat{s}_i))\overline{t}_i(1) + P(\hat{s}_i)\overline{t}_i(0) = (1 - P(\hat{s}_j))\overline{t}_j(1) + P(\hat{s}_j)\overline{t}_j(0) \\ \Leftrightarrow &\overline{t}_i(0) - (1 - P(\hat{s}_i))v_i(\hat{s}_i) = \overline{t}_j(0) - (1 - P(\hat{s}_j))v_j(\hat{s}_j)\end{aligned}$$

We can then write the principal's objective in terms of the transfer given for the action zero on one of the dimensions, the threshold in that dimension and the predictable utilities for the principal on all dimensions.

$$\begin{aligned}
W(t,g) &= \int_0^1 \sum_{a\in A}[-t(a) + \sum_{j\in N} w_j\, a_j\,]g(a|s)\,p(s)ds \\
&= -\int_0^1 \sum_{a\in A} t(a)g(a|s)\,p(s)ds + \sum_{j\in N} w_j \int_0^1 a_j\, g_j\,(a|s)\,p(s)ds \\
&= -\overline{t}_i(0) + (1-P(\hat{s}_i))v_i(\hat{s}_i) + \sum_{j\in N} w_j\,(1-P(\hat{s}_j\,)) \\
&= -\overline{t}_i(0) + \int_{\hat{s}_i}^1 (v_i(s) - \frac{1-P(s)}{p(s)} v_i'(s))\,p(s)ds + \sum_{j\in N} w_j\,(1-P(\hat{s}_j\,))
\end{aligned}$$

Step 2: We obtain an upper bound to the full problem by applying the minimax theorem. Any strategy g that is NIC must have a threshold form. Any threshold strategy can be NIC for beliefs given by Lemma 2 and from Proposition 2 we can construct a transfer function t that implements these beliefs.

The principal wants to maximize $W(t,g)$ and must implement beliefs that satisfy the statistical correctness constraints. The Lagrangian of the problem for maximizing this objective given this constraint can be written as follows, remembering that $\phi_i(s) = v_i(s) - \frac{1-P(s)}{p(s)} v_i'(s)$ and denoting Lagrange multipliers by $\lambda_j \in \mathbb{R}$ for the j th of the $n-1$ statistical correctness constraints. As any NIC strategy g is a threshold strategy inducing thresholds $\hat{s}$, we write the value of the Langrangian in terms of the vector of thresholds and expected transfers to the type choosing action 0 on any dimension $\overline{t}(0) = (\overline{t}_i(0))_{i\in N}$.

$$\begin{aligned}
&\overline{W}(\hat{s}, \overline{t}(0), \lambda) \\
&= -\overline{t}_i(0) + \int_{\hat{s}_i}^1 \phi_i(s)\,p(s)ds + \sum_{k\in N} w_k(1-P(\hat{s}_k)) \\
&+ \sum_{j\in N\setminus\{i\}} \lambda_j\,[-\overline{t}_j\,(0) + \int_{\hat{s}_j}^1 \phi_j\,(s)\,p(s)ds + \overline{t}_i(0) - \int_{\hat{s}_i}^1 \phi_i(s)\,p(s)ds] \\
&= (1 - \sum_{j\in N\setminus\{i\}} \lambda_j\,)[-\overline{t}_i(0) + \int_{\hat{s}_i}^1 \phi_i(s)\,p(s)ds] \\
&+ \sum_{j\in N\setminus\{i\}} \lambda_j\,[-\overline{t}_j\,(0) + \int_{\hat{s}_j}^1 \phi_j\,(s)\,p(s)ds] + \sum_{k\in N} w_k(1-P(\hat{s}_k))
\end{aligned}$$

Note that we can write the $n-1$ lagrange multipliers as $\beta = (\beta_1, ..., \beta_n) \in \mathbb{R}^n$ with $\sum_{j\in N} \beta_j = 1$ by setting $\beta_j = \lambda_j$ for $j \in N \setminus \{i\}$ and $\beta_i = 1 - \sum_{j\in N\setminus\{i\}} \lambda_j$.

The narrow participation constraints require that $\overline{t}_i(0) \geq 0$ for all $i \in N$.

Under a threshold strategy this reduces to the requirement that $\overline{t}_i(0) \geq 0$ for all $i \in N$. We can now write the principal's problem as follows

$$\sup_{\hat{s} \in [0,1]^n, \overline{t}(0) \in \mathbb{R}^n} \min_{\beta \in \mathbb{R}^n, \sum_{j \in N} \beta_j = 1} \overline{W}(\hat{s}, \overline{t}(0), \beta)$$

$$\text{subject to } \overline{t}_j(0) \geq 0 \text{ for } j \in N$$

Restricting the domain of β so that $\beta_j \in [0, 1]$ for all $j \in N$ in the minimization problem gives us the following upper bound.

$$\begin{aligned} &\min_{\tilde{\beta} \in [0,1]^n : \sum_{i \in N} \tilde{\beta}_i = 1} \sup_{\hat{s} \in [0,1]^n, \overline{t}(0) \in \mathbb{R}^n_{\geq 0}} \overline{W}(\hat{s}, \overline{t}(0), \beta) \\ \geq &\sup_{\hat{s} \in [0,1]^n, \overline{t}(0) \in \mathbb{R}^n_{\geq 0}} \min_{\tilde{\beta} \in [0,1]^n : \sum_{i \in N} \tilde{\beta}_i = 1} \overline{W}(\hat{s}, \overline{t}(0), \beta) \\ \geq &\sup_{\hat{s} \in [0,1]^n, \overline{t}(0) \in \mathbb{R}^n_{\geq 0}} \min_{\tilde{\beta} \in \mathbb{R}^n : \sum_{i \in N} \tilde{\beta}_i = 1} \overline{W}(\hat{s}, \overline{t}(0), \beta) \end{aligned}$$

We show that the objective function in our upper bound problem satisfies the conditions of the minimax theorem. This allows us to interchange the min and max operator and means we have a saddle point solution.

Define the quantile function $P^{-1}(s)$. Since $P(s)$ is strictly increasing, this is just the inverse and is also strictly increasing. For any vector $x \in [0, 1]^n$, we can write $P^{-1}(x_i) = \hat{s}_i$. Letting $P^{-1}(x) = (P^{-1}(x_i))_{i \in N}$, we use this to rewrite the objective.

$$\overline{W}(P^{-1}(x), \overline{t}(0), \beta) = \sum_{j \in N} [-\beta_j \overline{t}_j(0) + \int_{x_j}^{1} (\beta_j \phi_j(P^{-1}(u)) + w_j) du]$$

Taking derivatives of $\int_{x_j}^{1} (\beta_j \phi_j(P^{-1}(u)) + w_j) du$ with respect to the threshold x_j gives

$$-(\beta_j \phi_j(P^{-1}(x_j)) + w_j)$$

By the IVV assumption, this is decreasing and thus $\int_{x_j}^{1} (\beta_j \phi_j(P^{-1}(u)) + w_j) du$ is concave in $x \in [0, 1]^n$. We have that $-\beta_j \overline{t}_j(0)$ is also concave in $x \in [0, 1]^n$. The sum of concave functions is also concave, thus for fixed β, $\overline{W}(P^{-1}(x), \overline{t}(0), \beta)$ is concave in $\overline{t}(0)$ and the quantiles x. Since $\overline{W}(P^{-1}(x), \overline{t}(0), \beta)$ is convex in β for fixed $(\overline{t}(0), x)$ we can apply the minimax theorem (Sion, 1958) and obtain that

$$\begin{aligned}
&\min_{\beta\in[0,1]^n,\sum_{i\in N}\beta_i=1}\ \sup_{x\in[0,1]^n,\bar{t}(0)\in\mathbb{R}^n_{\geq 0}} \overline{W}(P^{-1}(x),\bar{t}(0),\beta)\\
&=\sup_{x\in[0,1]^n,\bar{t}(0)\in\mathbb{R}^n_{\geq 0}} \overline{W}(P^{-1}(x),\bar{t}(0),\beta^*)\\
&=\sup_{x\in[0,1]^n,\bar{t}(0)\in\mathbb{R}^n_{\geq 0}}\ \min_{\beta\in[0,1]^n,\sum_{i\in N}\beta_i=1} \overline{W}(P^{-1}(x),\bar{t}(0),\beta)
\end{aligned}$$

where β^* is the minimizer. We can then rewrite this in terms of thresholds using $\hat{s}_i = P^{-1}(x_i)$.

Step 3: We can show that we can attain the solution to this upper bound minimax problem with a mechanism that solves the full problem. Take the minimizing β^*. If $\beta^*_i > 0$ we must have that when comparing the dimension i and any other dimension j that

$$\begin{aligned}
&-\bar{t}_i(0)+\int_{\hat{s}_i}^1 (v_i(s)-\frac{1-P(s)}{p(s)}v_i'(s))\,p(s)\,ds\\
\leq &-\bar{t}_j(0)+\int_{\hat{s}_j}^1 (v_j(s)-\frac{1-P(s)}{p(s)}v_j'(s))\,p(s)\,ds
\end{aligned}$$

So that the dimension i term is minimal. At the solution the dimension i participation constraint binds $\bar{t}_i(0) = 0$ as otherwise the principal could increase the value of the objective by reducing $\bar{t}_i(0)$. If $\beta^*_j > 0$, then we have that this inequality must hold with equality with $\bar{t}_j(0) = 0$, which means the statistical correctness and participation constraint hold. If $\beta^*_j = 0$, then we can increase the value of $\bar{t}_j(0)$ without affecting the principal's objective value. From the above inequality, this can be done so that the statistical correctness constraint holds without violating the participation constraint. As $\beta^*_i > 0$, we have that $\bar{t}_i(0) = 0$ so the equality can be achieved with bounded $\bar{t}(0)$. Therefore we can replace the supremum with a maximum in the saddle point problem. The beliefs outlined then achieve the upper bound in the full problem, given that by Proposition 2 we can find a transfer function inducing these beliefs.

Proof of Proposition 3

For any fixed threshold strategy and fixed β the difference in the principal's objective can be written as

$$
\begin{aligned}
& W(\hat{s}) - \overline{W}(\hat{s};\beta) \\
&= \sum_{i\in N} \int_{\hat{s}_i}^{1} [(\phi_i(s) + w_i) - (\beta_i \phi_i(s) + w_i)]\, p(s)\, ds \\
&= \sum_{i\in N} (1-\beta_i) \int_{\hat{s}_i}^{1} (v_i(s) - \frac{1-P(s)}{p(s)} v_i'(s))\, p(s)\, ds \\
&= \sum_{i\in N} (1-\beta_i) v_i(\hat{s}_i)(1-P(\hat{s}_i))
\end{aligned}
$$

where the last line follows from the fact that $\int_{\hat{s}_i}^{1}(v_i(s) - \frac{1-P(s)}{p(s)} v_i'(s))\, p(s)\, ds = v_i(\hat{s}_i)(1-P(\hat{s}_i))$. Then as $\hat{s}_i \in [0,1]$ for the first case where $v_i(s) \leq 0$ clearly we have $W(\hat{s}) \leq \overline{W}(\hat{s};\beta)$ and for the second case where $0 \leq v_i(s)$ we have $W(\hat{s}) \geq \overline{W}(\hat{s};\beta)$.

Proof of Proposition 4

For every $i \in N$, let $\hat{s}_i^{rational}$ be the threshold such that $g_i(1|s) = \mathbb{1}\{s \geq \hat{s}_i^{rational}\}$ is the strategy that solves the principal's problem in the rational benchmark. Let $\hat{s}_i^{narrow}$ be the solution to the principal's problem with a narrow agent, as given by the solution to the minimax problem in Theorem 1. Let β^* be the saddle point shrinkage factors over dimensions from that problem.

When $v_i(s) \leq 0$ for all $s \in [0,1]$ we have $\phi_i(s) = v_i(s) - \frac{1-P(s)}{p(s)} v_i'(s) \leq 0$ for all $s \in [0,1]$, then $\phi_i(s) + w_i \leq \beta_i^* \phi_i(s) + w_i$. Suppose that for some dimension $j \in N$ we have $\hat{s}_j^{narrow} > \hat{s}_j^{rational}$. Optimality of $\hat{s}_j^{rational}$ as a threshold implies $\phi_j(s) + w_j > 0$ for all $s \in (\hat{s}_j^{rational}, 1]$. However, then $\hat{s}_j^{narrow} > \hat{s}_j^{rational}$ cannot be optimal for the principal as $0 < \phi_j(s) + w_j \leq \beta_j^* \phi_j(s) + w_j$ for $s \in (\hat{s}_j^{rational}, \hat{s}_j^{narrow}]$, a contradiction.

For the second case we again assume that $\beta_i^* > 0$, as otherwise $\hat{s}_i^{narrow} = 1$ in which case the result holds. Then $\phi_i(s) = v_i(s) - \frac{1-P(s)}{p(s)} v_i'(s) \geq 0$ for any $s \in [\hat{s}_i^{narrow}, 1]$. Otherwise there is an $\tilde{s}_i \in (\hat{s}_i^{narrow}, 1]$ such that by IVV $\beta_i^* \phi_i(s) + w_i < 0$ for all $s \in [\hat{s}_i^{narrow}, \tilde{s}_i]$, and the principal could then switch to implementing $g_i(0|s) = 1$ for all $s \in [\hat{s}_i^{narrow}, \tilde{s}_i]$ and obtain a higher payoff.

Now for contradiction assume $\hat{s}_i^{rational} > \hat{s}_i^{narrow}$. We have that $\phi_i(s) \geq 0$ and thus $\phi_i(s) + w_i \geq \beta_i^* \phi_i(s) + w_i$ for any $s \in [\hat{s}_i^{narrow}, \hat{s}_i^{rational}]$. Then optimality of $\hat{s}_i^{rational}$ is contradicted by the fact that optimality of $\hat{s}_i^{narrow}$ requires $0 < \beta_i^* \phi_i(s) + w_i \leq \phi_i(s) + w_i$ for all $s \in (\hat{s}_i^{narrow}, \hat{s}_i^{rational}]$, where the first strict inequality follows from IVV.

Proof of Proposition 5

For the first case, for fixed thresholds $(\hat{s}_i)_{i \in N}$ we can take the expression for the principal's welfare in Theorem 1 and show that moving from N to split M results in weakly higher welfare for the principal.

$$
\begin{aligned}
&\min_{\beta \in [0,1]^n : \sum_{i \in N} \beta_i = 1} \overline{W}(\hat{s}; \beta : N) \\
&= \min_{\beta \in [0,1]^n : \sum_{i \in N} \beta_i = 1} \sum_{i \in N} \beta_i v_i(\hat{s}_i)(1 - P(\hat{s}_i)) + \sum_{i \in N} w_i (1 - P(\hat{s}_i)) \\
&= \min_{\beta \in [0,1]^n : \sum_{i \in N} \beta_i = 1} \sum_{i \in N} \beta_i (1 - P(\hat{s}_i)) \sum_{j \in \mathcal{P}(i)} v_j(\hat{s}_j) + \sum_{i \in N} (1 - P(\hat{s}_i)) \sum_{j \in \mathcal{P}(i)} w_j \\
&\leq \min_{\beta \in [0,1]^m : \sum_{j \in M} \beta_j = 1} \sum_{i \in N} (1 - P(\hat{s}_i)) \sum_{j \in \mathcal{P}(i)} \beta_j v_j(\hat{s}_j) + \sum_{i \in N} (1 - P(\hat{s}_i)) \sum_{j \in \mathcal{P}(i)} w_j \\
&= \min_{\beta \in [0,1]^m : \sum_{j \in M} \beta_j = 1} \overline{W}(\hat{s}; \beta : M)
\end{aligned}
$$

Where the inequality holds as in this case for any $i \in N$ and $j \in \mathcal{P}(i)$, $0 \geq v_j(s) \geq v_i(s)$ for all $s \in S$.

For the second case, the principal also faces a loss under the merged dimension space from a reduction in the space of threshold strategies that can be implemented. This is because under the merged space for any given dimension $i \in N$, all the thresholds $\hat{s}_j$, $j \in \mathcal{P}(i)$ must be equalized.

Given a fixed threshold strategy in the split space $(\hat{s}_j)_{j \in M}$, define a threshold strategy in the merged space such that $\hat{s}_i = \max_{j \in \mathcal{P}(i)} \hat{s}_j$. For any $j(i) = argmax_{k \in \mathcal{P}(i)} \hat{s}_k$ we can focus on solutions to the split space minimax problem in Theorem 1 with $\beta^*_{j(i)} > 0$. If $\beta^*_{j(i)} = 0$ it would be optimal for the principal to implement the threshold $\hat{s}_{j(i)} = 1$ because $w_{j(i)} < 0$. If $\hat{s}_j = 1$ for any dimension, then as $v_k(s)(1 - P(s)) > 0$ for all $k \in M$[6] and $s \in (0, 1)$ at the solution to the minimax problem we can only have $\beta^*_h > 0$ for dimensions with $\hat{s}_h = 1$. Then the principal's payoff is zero under the split space, which gives the result as the principal can also obtain a zero payoff under the merged space.

With $\beta^*_{j(i)} > 0$ for every $i \in N$, we can show that the principal benefits from

[6] As $v_k(s)$ is strictly increasing in s and $v_k(0) \geq 0$.

moving from the threshold strategy $(\hat{s}_j)_{j \in M}$ in the split space to the threshold strategy with $\hat{s}_i = \max_{j \in \mathcal{P}(i)} \hat{s}_j$ for all $i \in N$ in the merged space. This then gives the result.

$$\begin{aligned}
&\min_{\beta \in [0,1]^m : \sum_{i \in M} \beta_j = 1} \max_{(\hat{s}_j)_{j \in M} \in [0,1]^m} \overline{W}(\hat{s}; \beta : M) \\
&= \min_{k \in N} v_{j(k)}(\hat{s}^*_{j(k)})(1 - P(\hat{s}^*_{j(k)})) + \sum_{j \in M} w_j (1 - P(s^*_j)) \\
&\leq \min_{k \in N} \sum_{h \in \mathcal{P}(k)} v_h(\hat{s}^*_{j(k)})(1 - P(\hat{s}^*_{j(k)})) + \sum_{i \in N} (1 - P(\hat{s}^*_{j(i)})) \sum_{j \in \mathcal{P}(i)} w_j \\
&= \min_{\beta \in [0,1]^n : \sum_{i \in N} \beta_i = 1} \sum_{i \in N} \beta_i v_i(\hat{s}_{j(i)})(1 - P(\hat{s}_{j(i)})) + \sum_{i \in N} w_i (1 - P(\hat{s}_{j(i)})) \\
&\leq \min_{\beta \in [0,1]^n : \sum_{i \in N} \beta_i = 1} \max_{(\hat{s}_i)_{i \in N} \in [0,1]^n} \overline{W}(\hat{s}; \beta : N)
\end{aligned}$$

The first equality holds because $\beta^*_{j(k)} > 0$ means that $j(k) = \arg\min_{i \in M} v_i(\hat{s}_{j(k)})(1 - P(\hat{s}_{j(k)}))$. The first inequality holds because $v_h(s) \geq 0$ for all $s \in S$ and $\hat{s}_{j(i)} \geq \hat{s}_k$ for all $k \in \mathcal{P}(i)$, which means the weight attached to any $w_k < 0$ is lower.

Proof of Lemma 3

From Proposition 1, it is clear that under symmetry with a rational agent the principal wants to equalize thresholds.

For the narrow case, let $(\hat{s}^*, \beta^*)$ be the saddle point solution to the characterization problem in Theorem 1. If for any pair of dimensions $i, j \in N$, $\beta^*_i, \beta^*_j \in (0, 1)$ then

$$\int_{\hat{s}_i}^{1} \left(v(s) - \frac{1 - P(s)}{p(s)} v'(s)\right) p(s)\, ds = \int_{\hat{s}_j}^{1} \left(v(s) - \frac{1 - P(s)}{p(s)} v'(s)\right) p(s)\, ds$$

as otherwise we would have $\beta^*_k = 0$ for one of the dimensions $k = \{i, j\}$ as putting any weight on that dimension would not be minimizing.

The remaining case is when $\beta^*_i = 0$ for some $i \in N$. Then it is optimal for the principal to implement threshold $\hat{s}^*_i = 0$ when $w > 0$ and $\hat{s}^*_i = 1$ when $w < 0$. But then in both cases either all dimensions either choose the same marginal strategy as on i, in which case the result holds, or the solution to the minimization problem would be to have $\beta^*_i > 0$, a contradiction. To see this, in the first case for any

$j \in N$ unless $\hat{s}_j^* = 0$ we have.

$$\begin{aligned}
&\int_0^1 (v(s) - \frac{1-P(s)}{p(s)} v'(s))\, p(s) ds \\
&< \int_{\hat{s}_j^*}^1 (v(s) - \frac{1-P(s)}{p(s)} v'(s))\, p(s) ds = v(\hat{s}_j^*)(1-P(\hat{s}_j^*))
\end{aligned}$$

as $v(s) - \frac{1-P(s)}{p(s)} v'(s) < 0$ for all $s \in [0, \hat{s}_j^*)$, as otherwise under IVV and $w > 0$ it would be optimal to set $\hat{s}_j^* = 0$. For the second case, for all $j \in N$ unless $\hat{s}_j^* = 1$

$$\begin{aligned}
0 &= \int_1^1 (v(s) - \frac{1-P(s)}{p(s)} v'(s))\, p(s) ds \\
&< \int_{\hat{s}_j^*}^1 (v(s) - \frac{1-P(s)}{p(s)} v'(s))\, p(s) ds = v(\hat{s}_j^*)(1-P(\hat{s}_j^*))
\end{aligned}$$

as $v(s) - \frac{1-P(s)}{p(s)} v'(s) > 0$ for all $s \in (\hat{s}_j^*, 1]$, as otherwise since $w < 0$ then under IVV $\hat{s}_j^* = 1$ would be optimal.

We see that we must have that $\beta_i^* \in (0,1)$ for all $i \in N$. Rearranging the implied equality, we have that for any $i, j \in N$

$$\begin{aligned}
&\int_{\hat{s}_i^*}^1 (v(s) - \frac{1-P(s)}{p(s)} v'(s))\, p(s) ds = \int_{\hat{s}_j^*}^1 (v(s) - \frac{1-P(s)}{p(s)} v'(s))\, p(s) ds \\
\Leftrightarrow &\int_{\hat{s}_i^*}^{\hat{s}_j^*} (v(s) - \frac{1-P(s)}{p(s)} v'(s))\, p(s) ds = 0
\end{aligned}$$

which implies $\hat{s}_i^* = \hat{s}_j^*$ by IVV.

Proof of Proposition 6

For any fixed n, by Lemma 3 we have that $\hat{s}_i^{(n)} = \hat{s}^{(n)}$ for all $i \in N$ under the symmetric dimension space assumption. Given the principal's optimal thresholds $\hat{s}^{(n)}$, the saddle point problem in Theorem 1 is

$$\begin{aligned}
&\min_{\beta \in [0,1]^n, \sum_{i \in N} \beta_i = 1} \overline{W}(\hat{s}^{(n)}, \beta) \\
&= \sum_{i \in N} [\beta_i \int_{\hat{s}^{(n)}}^1 (\frac{1}{n} v(s) - \frac{1-P(s)}{p(s)} \frac{1}{n} v'(s))\, p(s) ds + \frac{1}{n} w(1 - P(\hat{s}^{(n)}))]
\end{aligned}$$

This has solution $\beta_i = \frac{1}{n}$ for all $i \in N$. The optimal threshold $\hat{s}^n$ can then be characterized by the following variational inequality. For all $s \in S$ we must have

$$(s - \hat{s}^{(n)})(\frac{1}{n}(\frac{1}{n}v(s) - \frac{1 - P(s)}{p(s)}\frac{1}{n}v'(s)) + \frac{1}{n}w) \geq 0$$

$$\Leftrightarrow$$

$$(s - \hat{s}^{(n)})((\frac{1}{n}v(s) - \frac{1 - P(s)}{p(s)}\frac{1}{n}v'(s)) + w) \geq 0$$

If $w > 0$, there exists an $\underline{n}$ such that for all $\tilde{n} \geq \underline{n}$[7]

$$\frac{1}{\tilde{n}}(v(0) - \frac{1 - P(0)}{p(0)}v'(0)) + w > 0$$

By IVV, we then have that $(\frac{1}{n}v(s) - \frac{1-P(s)}{p(s)}\frac{1}{n}v'(s)) + w > 0$ for all $s \in S$. Thus the only solution to the variational inequality is $\hat{s}^{(\tilde{n})} = 0$, as if $\hat{s}^{(n)} > 0$ then the inequality is violated for all $s \in [0, \hat{s}^{(n)}]$.

We can make an analogous argument when $w < 0$ to show the second part. For the final part, we can implement the symmetric threshold strategy with transfers $t(1, ..., 1) = \frac{1}{n}v(\hat{s}^{(n)})$ and $t(0, ..., 0) = 0$ (only action combinations $(1, ..., 1)$ and $(0, ..., 0)$ are chosen with positive probability). Thus $g(1, ..., 1)t(1, ..., 1) + g(0, ..., 0)t(0, ..., 0) \to 0$ as $n \to \infty$.

Proof of Proposition 7

Proof. Take any solution that attains the Principal's maximal value under Theorem 1; $(\hat{s}', \beta')$. Define the minimal value of the transfer component under this threshold strategy; $\phi_{\text{min}} = \min_{i \in N} v_i(\hat{s}'_i)(1 - P(\hat{s}'_i))$. We modify the solution as follows: for any $j \in N$ such that $w_j = 0$ and $v_i(\hat{s}'_j)(1 - P(\hat{s}'_j)) > \phi_{\text{min}}$, if $\phi_{\text{min}} \leq 0$ set $\hat{s}^*_j = 1$ and if $\phi_{\text{min}} > 0$ set $\hat{s}^*_j$ such that $v_i(\hat{s}^*_j)(1 - P(\hat{s}^*_j)) = \phi_{\text{min}}$.[8] Set $\beta^* = \beta'$ and $\hat{s}^*_i = \hat{s}'_i$ for all other dimensions. The new solution $(\hat{s}^*, \beta^*)$ achieves the same value to the principal as the original solution, since $v_i(\hat{s}^*_i)(1 - P(\hat{s}^*_i)) \geq \phi_{min}$ for all $i \in N$ and $\hat{s}^*_i = \hat{s}'_i$ if $w_i \neq 0$. The new solution $(\hat{s}^*, \beta^*)$ therefore solves the saddle point problem of Theorem 1.

[7] Since $v(s)$ is bounded, as it is continuously differentiable with domain $[0, 1]$.

[8] Such an $\hat{s}^*_j \in [\hat{s}'_j, 1]$ must exist by the intermediate value theorem since $v_i(\hat{s}'_j)(1-P(\hat{s}'_j)) > \phi_{\text{min}} > 0$, $v_j(1)(1 - P(1)) = 0$ and $v_j(\hat{s}_j)(1 - P(\hat{s}_j))$ is continuous in $\hat{s}_j \in [\hat{s}'_j, 1]$.

Given $(\hat{s}^*, \beta^*)$, denote the following minimizing set of dimensions

$$\underline{N} \equiv \arg\min_{i \in N} v_i(\hat{s}_i^*)(1 - P(\hat{s}_i^*))$$

We can then denote the subset of the greatest dimensions in $\underline{N}$ by

$$\underline{N}^{max} \equiv \arg\max_{j \in \underline{N}} \hat{s}_j^*$$

Choose some $i^* \in \underline{N}^{max}$, and define the transfer function as in the statement of the proposition. First consider the case that $\hat{s}_{i^*}^* = \max_{i \in \underline{N}} \hat{s}_i^* = 1$. Then we have that $\arg\min_{i \in N} v_i(\hat{s}_i^*)(1 - P(\hat{s}_i^*)) = 0$, and all $j \in \underline{N}$ must be such that either $\hat{s}_j^* = 1$ or $v_j(\hat{s}_j^*) = 0$. Any non-minimizing dimension $k \in N \setminus \{\underline{N}\}$ must have $\hat{s}_k^* = 0$ and $w_k > 0$. This is because by the modification at the start of the proof, any non-minimizing $k \in N \setminus \{\underline{N}\}$ must be such that $\hat{s}_k^* \in \{0, 1\}$ and $\hat{s}_k^* \neq 1$ otherwise k would be in the set of minimizing dimensions $\underline{N}$. Since $\arg\min_{i \in N} v_i(\hat{s}_i^*)(1 - P(\hat{s}_i^*)) = 0$ it must be the case that $v_k(\hat{s}_k^* = 0) \geq 0$. All this means that this threshold strategy is NIC under beliefs where for all $i \in N$ $\overline{t}_i(1) = \overline{t}_i(0) = 0$. These beliefs clearly also satisfy the narrow participation and statistical correctness constraints, and hold under the transfer function we have defined as there is no action where a_{i^*} is chosen with positive probability under this threshold strategy, as $\hat{s}_{i^*}^* = 1$.

Now consider the case where $\hat{s}_{i^*}^* = \max_{i \in \underline{N}} \hat{s}_i^* < 1$. We show that for each dimension $j \in N$, there are beliefs consistent with the proposed transfer function and threshold strategy $\hat{s}^*$ such that NIC and narrow participation holds for that dimension. We also show that the expected beliefs for any dimension must be equal to the same value; $P(\hat{s}_j^*)\overline{t}_j(0) + (1 - P(\hat{s}_j^*))\overline{t}_j(1) = -v_{i^*}(\hat{s}_{i^*}^*)(1 - P(\hat{s}_{i^*}^*))$, which gives us that the statistical correctness constraint holds. Together this means we have our result.

For any dimension $j \in N$ with an interior threshold; $\hat{s}_j^* \in (0, 1)$, it must be the case that $j \in \underline{N}$ as the modification at the start of the proof rules out non-minimizing interior thresholds. Since $i^* \in \underline{N}^{max}$, $\hat{s}_{i^*}^*$ is a maximal interior threshold. This means that there is no action combination $\tilde{a} \in A$ such that $\tilde{a}_{i^*} = 1$ and $\tilde{a}_j = 0$ that has positive probability under the threshold strategy; $g(\tilde{a}) > 0$. Since only action combinations $a \in A$ where $a_{i^*} = 1$ have $t(a) \neq 0$ under the transfer function we have outlined, this means that the beliefs of the narrow agent on dimension j are such that $\overline{t}_j(0) = 0$ and

$$\overline{t}_j(1) = \sum_{a_{-j} \in A_{-j}} t(a_j, a_{-j}) \frac{g(a_j, a_{-j})}{g_j(a_j)} = -v_{i^*}(\hat{s}^*_{i^*}) \frac{1 - P(\hat{s}^*_{i^*})}{1 - P(\hat{s}^*_j)}$$

These beliefs satisfy NIC for this dimension as because i^*, $j \in \underline{N}^{max}$ we have that $\overline{t}_j(1) - \overline{t}_j(0) = -v_{i^*}(\hat{s}^*_{i^*}) \frac{1-P(\hat{s}^*_{i^*})}{1-P(\hat{s}^*_j)} = -v_j(\hat{s}^*_j)$. Narrow participation is also satisfied as

$$\overline{U}_j(s) = \mathbb{1}[s \geq \hat{s}^*_j](v_j(s) - v_j(\hat{s}^*_j)) \geq 0$$

The expected belief on this dimension over all types is then

$$P(\hat{s}^*_j)\overline{t}_j(0) + (1 - P(\hat{s}^*_j))\overline{t}_j(1) = -v_j(\hat{s}^*_j)(1 - P(\hat{s}^*_j)) = -v_{i^*}(\hat{s}^*_{i^*})(1 - P(\hat{s}^*_{i^*}))$$

When $j \in N$ is such that $\hat{s}^*_j = 1$, we must have that j is non-minimizing as $\max_{i \in \underline{N}} \hat{s}^*_i < 1$. Since all types choose $a_j = 0$ with probability one on this dimension, we have that $\overline{t}_j(0) = -v_{i^*}(\hat{s}^*_{i^*})(1 - P(\hat{s}^*_{i^*}))$. Narrow participation then holds as since $j \notin \underline{N}$, $\overline{t}_j(0) = -v_{i^*}(\hat{s}^*_{i^*})(1 - P(\hat{s}^*_{i^*})) \geq -v_j(\hat{s}^*_j)(1 - P(\hat{s}^*_j)) = 0$. We have that the expected belief is $\overline{t}_j(0) = -v_{i^*}(\hat{s}^*_{i^*})(1 - P(\hat{s}^*_{i^*}))$, and since $a_j = 1$ is off-path we can set $\overline{t}_j(1) = \overline{t}_j(0) - v_j(1)$ so that NIC holds.

Finally, when $j \in N$ has zero threshold $\hat{s}^*_j = 0$ then for the same reasons as in the interior case we have that $\overline{t}_j(1) = -v_{i^*}(\hat{s}^*_{i^*}) \frac{1-P(\hat{s}^*_{i^*})}{1-P(\hat{s}^*_j)} = -v_{i^*}(\hat{s}^*_{i^*})(1 - P(\hat{s}^*_{i^*}))$. This is also equal to the expected belief for this dimension. Narrow participation holds as because i^* is a minimal dimension $i^* \in \underline{N}$ we have $v_j(0) \geq v_{i^*}(\hat{s}^*_{i^*})(1 - P(\hat{s}^*_{i^*}))$ and therefore.

$$\overline{U}_j(s) = (v_j(s) - v_{i^*}(\hat{s}^*_{i^*})(1 - P(\hat{s}^*_{i^*}))) \geq 0$$

Again, because $a_j = 0$ is off path we can set $\overline{t}_j(0) = \overline{t}_j(1) + v_j(0)$ to ensure NIC. □

References

[1] Armstrong, M. (2016). Nonlinear pricing. *Annual Review of Economics 8*(1), 583–614. 21

[2] Bohren, J. A. and D. N. Hauser (2021). Learning with heterogeneous misspecified models: Characterization and robustness. *Econometrica 89*(6), 3025–3077. 8

[3] Camara, M. K. (2022). Computationally tractable choice. In *EC*, pp. 28. 8

[4] Carroll, G. (2017). Robustness and separation in multidimensional screening. *Econometrica 85*(2), 453–488. 5, 13, 14, Online Appendix: 3

[5] Eisenhuth, R. (2019). Reference-dependent mechanism design. *Economic Theory Bulletin 7*(1), 77–103. 9

[6] Eliaz, K. and R. Spiegler (2006). Contracting with diversely naive agents. *The Review of Economic Studies 73*(3), 689–714. 9

[7] Eliaz, K. and R. Spiegler (2020). A model of competing narratives. *American Economic Review 110*(12), 3786–3816. 8

[8] Eliaz, K. and R. Spiegler (2024). News media as suppliers of narratives (and information). *arXiv preprint arXiv:2403.09155*. 8

[9] Enke, B. and F. Zimmermann (2019). Correlation neglect in belief formation. *The review of economic studies 86*(1), 313–332. 7

[10] Esponda, I. and D. Pouzo (2016). Berk–nash equilibrium: A framework for modeling agents with misspecified models. *Econometrica 84*(3), 1093–1130. 8

[11] Eyster, E. and M. Rabin (2005). Cursed equilibrium. *Econometrica 73*(5), 1623–1672. 9

[12] Farhi, E. and X. Gabaix (2020). Optimal taxation with behavioral agents. *American Economic Review 110*(1), 298–336. 9

[13] Fernbach, P. M., A. Darlow, and S. A. Sloman (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science 21*(3), 329–336. 8

[14] Frick, M., R. Iijima, and Y. Ishii (2020). Misinterpreting others and the fragility of social learning. *Econometrica 88*(6), 2281–2328. 8

[15] Fudenberg, D., G. Lanzani, and P. Strack (2021). Limit points of endogenous misspecified learning. *Econometrica 89*(3), 1065–1098. 8

[16] Graeber, T. (2023). Inattentive inference. *Journal of the European Economic Association 21*(2), 560–592. 8

[17] He, S. and S. Kučinskas (2024). Expectation formation with correlated variables. *The Economic Journal 134*(660), 1517–1544. 8

[18] Heidhues, P. and B. Kőszegi (2010). Exploiting naivete about self-control in the credit market. *American Economic Review 100*(5), 2279–2303. 9

[19] Heidhues, P., B. Kőszegi, and P. Strack (2018). Unrealistic expectations and misguided learning. *Econometrica 86*(4), 1159–1214. 8

[20] Herweg, F., D. Müller, and P. Weinschenk (2010). Binary payment schemes: Moral hazard and loss aversion. *American Economic Review 100*(5), 2451–2477. 9

[21] Jehiel, P. (2005). Analogy-based expectation equilibrium. *Journal of Economic Theory 123*(2), 81–104. 9, Online Appendix: 1

[22] Jehiel, P. (2011). Manipulative auction design. *Theoretical economics 6*(2), 185–217. 9

[23] Jehiel, P. and K. Mierendorff (2024). Auction design with data-driven misspecifications: Inefficiency in private value auctions with correlation. *Theoretical Economics*, (Forthcoming). 9

[24] Kőszegi, B. (2014). Behavioral contract theory. *Journal of Economic Literature 52*(4), 1075–1118. 9

[25] Kőszegi, B. and F. Matějka (2020). Choice simplification: A theory of mental budgeting and naive diversification. *The Quarterly Journal of Economics 135*(2), 1153–1207. 8

[26] Lian, C. (2021). A theory of narrow thinking. *The Review of Economic Studies 88*(5), 2344–2374. 8

[27] Lockwood, B. B. (2020). Optimal income taxation with present bias. *American Economic Journal: Economic Policy 12*(4), 298–327. 9

[28] Milgrom, P. and I. Segal (2002). Envelope theorems for arbitrary choice sets. *Econometrica 70*(2), 583–601. Online Appendix: 2

[29] Mirman, L. J. and D. Sibley (1980). Optimal nonlinear prices for multiproduct monopolies. *The Bell Journal of Economics 11*(2), 659–670. 14

[30] Myerson, R. B. (1981). Optimal auction design. *Mathematics of operations research 6*(1), 58–73. 19, Online Appendix: 12, Online Appendix: 14

[31] O'Donoghue, T. and M. Rabin (2006). Optimal sin taxes. *Journal of Public Economics 90*(10-11), 1825–1849. 9

[32] Pearl, J. (2009). *Causality*. Cambridge university press. 2

[33] Piccione, M. and A. Rubinstein (2003). Modeling the economic interaction of agents with diverse abilities to recognize equilibrium patterns. *Journal of the European economic association 1*(1), 212–223. 9

[34] Rabin, M. and G. Weizsäcker (2009). Narrow bracketing and dominated choices. *American Economic Review 99*(4), 1508–1543. 8

[35] Read, D., G. Loewenstein, M. Rabin, G. Keren, and D. Laibson (1999). Choice bracketing. *Journal of Risk and Uncertainty 19*(1/3), 171–197. 8

[36] Rubinstein, A. (1993). On price recognition and computational complexity in a monopolistic model. *Journal of Political Economy 101*(3), 473–484. 8

[37] Schumacher, H. and H. C. Thysen (2022). Equilibrium contracts and boundedly rational expectations. *Theoretical Economics 17*(1), 371–414. 8

[38] Sion, M. (1958). On general minimax theorems. *Pacific Journal of Mathematics 8*(1), 171 – 176. 35

[39] Spiegler, R. (2016). Bayesian networks and boundedly rational expectations. *The Quarterly Journal of Economics 131*(3), 1243–1290. 8

[40] Spiegler, R. (2020). Behavioral implications of causal misperceptions. *Annual Review of Economics 12*(1), 81–106. 17

[41] Spinnewijn, J. (2015). Unemployed but optimistic: Optimal insurance design with biased beliefs. *Journal of the European Economic Association 13*(1), 130–167. 9

[42] Thaler, R. (1985). Mental accounting and consumer choice. *Marketing science 4*(3), 199–214. 8

[43] Thaler, R. H. (1999). Mental accounting matters. *Journal of Behavioral decision making 12*(3), 183–206. 8

[44] Tversky, A. and D. Kahneman (1981). The framing of decisions and the psychology of choice. *science 211*(4481), 453–458. 8

[45] Vorjohann, P. (2024). Reference-dependent choice bracketing. *Available at SSRN: https://ssrn.com/abstract=4927177*. 8

Online Appendix

Alexander Clyde[‡‡]

August 20, 2025

A Connection to ABEE

It is possible to express behaviour under narrow inference as an Analogy Based Expectation Equilibrium (ABEE) (Jehiel, 2005). Under ABEE, each player in a game has an ‘analogy partition’ of the set of histories where other players move. For any cell in the partition, a player believes that the strategy of the other players is the average of the true strategies for histories in that cell.

Take a game with $n + 2$ players; consisting of the principal, n different ‘selves’ of the agent and a player of nature. Each of the n selves corresponds to one of the n actions available to the agent, so that self $i \in \{1, ..., n\}$ controls action a_i. All selves share identical preferences over the actions and transfer. The timing of the game is as follows; first the principal chooses a transfer function t. Then the common type of the agent's selves is drawn. After learning this common type then, moving in any order, each of the n selves choose either an action from the set they control or to not participate in the mechanism. Finally, the player of nature implements the transfer function chosen by the principal.

Although each of the agent's selves have common preferences, they differ according to their analogy partitions. Each self partitions the history at which the player of nature moves, with each cell in the partition corresponding to a different action chosen by the self. Thus their beliefs about the expected transfer from each action is the average transfer obtained among all types of agents choosing that action. This coincides with the beliefs under narrow inference. Behaviour under narrow inference then coincides with an ABEE of this multi-selves game.

[‡‡] Department of Economics, Aalto University School of Business; Address: Ekonominaukio 1, FI-02150 Espoo, Finland; Email: alexander.clyde.econ@gmail.com or alexander.clyde@aalto.fi

B Rational Benchmark

We analyze the principal's optimal mechanism in the rational benchmark. The next result provides a standard characterization of all IC strategies and expected utilities. We say that expected utilities $\{U(s)\}_{s\in S}$ can be *achieved* given strategy g if there exists a transfer function t such that for every type $s \in S$, $U(s)$ is the expected utility.

Lemma B.1. *A strategy g and expected utilities $\{U(s)\}_{s\in S}$ that can be achieved given g are IC only if*

1. *Cyclical monotonicity condition: For any subset of types $\{s_1, ..., s_{k+1}\} \subset S$ with $s_{k+1} = s_1$*

$$\sum_{m=1}^{k} \sum_{i\in N} (v_i(s_{m+1}) - v_i(s_m)) \sum_{a_i\in A_i} g_i(a_i|s_{m+1})a_i \geq 0 \tag{OA:2}$$

2. *The expected utility $U(s)$ is increasing in $s \in S$.*

3. *The following envelope condition holds for any two types $s, s' \in S$*

$$U(s) = U(s') + \sum_{i\in N} \int_{s'}^{s} v_i'(z) \sum_{a_i\in A_i} a_i g(a_i|z)dz \tag{OA:3}$$

Proof. For any types $s, s' \in S$, the rational incentive constraints (5) require that

$$U(s) \geq U(s') + \sum_{i\in N} (v_i(s) - v_i(s')) \sum_{a_i\in A_i} g_i(a_i|s')a_i$$

This then implies the second condition. Iterating these rewritten ICs along any cycle of types $\{s_1, ..., s_{k+1}\}$ with $s_{k+1} = s_1$ gives that the cyclical monotonicity condition must hold. We can also use the rewritten ICs to show the envelope condition holds by using the Lipschitz continuity arguments in Theorems 1 and 2 of Milgrom and Segal (2002). □

As noted, a strategy where the expected utility of actions in individual dimensions is non-monotonic in type can be IC as long as the cyclical monotonicity condition is satisfied. This is in contrast to the narrow agent model where the action has to be monotonic in type for each dimension. We shall see that under

IVV, this does not matter as the principal's optimal mechanism implements a threshold strategy that is monotone on each dimension anyway.

The example in Section 3.2 of Carroll (2017) shows that we can have non-separability in an optimal selling mechanism with a co-monotonic type distribution. This is due to the different monotonicity condition when we have multiple goods relative to when we have a single good. With a single good, we have that under IC higher types must get the good with higher probability, while with multiple goods we can trade-off probabilities across goods without violating IC. In the example of Section D, we show that when IVV does not hold this also applies in our model.

We use Lemma B.1 to prove the next lemma, showing that any threshold strategy can be implemented with an additively separable transfer function.

Lemma B.2. *Let g^* be a threshold strategy and $U(0)$ be the expected utility of the type $s = 0$. The strategy is IC and achieves the expected utility $U(0)$ for type $s = 0$ under transfer function*

$$t(a_1, \dots, a_n) = \sum_{i \in N} t^i(a_i) \tag{OA:4}$$

$$t^i(0) = \frac{1}{n} U(0), t^i(1) = -v_i(\hat{s}_i) + \frac{1}{n} U(0) \;\; \textit{for all } i \in N \tag{OA:5}$$

Proof. The expected utility of type s behaving according to threshold strategy g^* when the transfer function is that given in the lemma statement is

$$\begin{aligned} U(s) &= \sum_{i \in N} (v_i(s) + t^i(1) - t^i(0))\mathbb{1}[s \geq \hat{s}_i] + \sum_{i \in N} t^i(0) \\ &= \sum_{i \in N} (v_i(s) - v_i(\hat{s}_i))1[s \geq \hat{s}_i] + U(0) \end{aligned}$$

We can see from this that the threshold strategy is IC, as for any dimension $i \in N$ a type above the threshold $\hat{s}_i$ gets a weakly positive utility from choosing $a_i = 1$ over $a_i = 0$ while a type below the threshold gets a negative utility. The lowest type $s = 0$ gets utility $U(0)$. □

B.0.1 Proof of Proposition 1

Proof. We can rewrite the principal's objective (2) using the envelope formula (OA:3) from Lemma B.1 and the expression for the transfer function in terms of utilities in the direct mechanism.

$$
\begin{aligned}
W(t,g) &= \int_0^1 \sum_{a\in A}[-t(a) + \sum_{i\in N} w_i a_i] g(a|s)\, p(s)\, ds \\
&= \sum_{i\in N}\int_0^1 \sum_{a_i\in A_i}[a_i v_i(s) + w_i a_i] g_i(a_i|s)\, p(s)\, ds - \int_0^1 U(s)\, p(s)\, ds \\
&= \int_0^1 \sum_{a\in A}[\sum_{i\in N} a_i v_i(s) + \sum_{i\in N} w_i a_i] g(a|s)\, p(s)\, ds \\
&- \sum_{i\in N}\int_0^1 [\int_0^s v_i'(z) \sum_{a_i\in A_i} a_i g_i(a_i|z)\, dz]\, p(s)\, ds - U(0) \\
&= \int_0^1 \sum_{a\in A}[\sum_{i\in N} a_i v_i(s) + \sum_{i\in N} w_i a_i] g(a|s)\, p(s)\, ds \\
&- \sum_{i\in N}\int_0^1 [\int_z^1 p(s)\, ds] v_i'(z) \sum_{a_i\in A_i} a_i g_i(a_i|z)\, dz - U(0) \\
&= \sum_{i\in N}\int_0^1 \sum_{a_i\in A_i}[a_i v_i(s) + w_i a_i] g_i(a_i|s)\, p(s)\, ds \\
&- \sum_{i\in N}\int_0^1 [1-P(z)] v_i'(z) \sum_{a_i\in A_i} a_i g_i(a_i|z)\, dz - U(0) \\
&= \sum_{i\in N}\int_0^1 \sum_{a_i\in A_i}[\phi_i(s) a_i + w_i a_i] g_i(a_i|s)\, p(s)\, ds - U(0)
\end{aligned}
$$

Where the first line follows from expressing the transfer function in terms of expected utility and the last two lines follow from a standard switching of the order of integration and rewriting in terms of marginal strategies.

Clearly it is optimal to set the expected utility of the lowest type to zero. We now consider a relaxed version of the Principal's problem where we ignore the cyclical monotonicity constraints from Lemma B.1 and that expected utilities might not be achieved given g. We show that under IVV, the mechanism that solves this relaxed problem implements a threshold strategy. Under a threshold strategy we have that for all $i \in N$, $\sum_{a_i\in A_i} a_i g_i(a_i|s)$ is increasing in type s.

By Lemma B.2 we can find a transfer function that implements the threshold strategy as IC and achieves any given expected utility for the lowest type. Thus the solution to the relaxed problem coincides with the solution to the full problem.

$$
\max_{g_i\in\Delta(A_i)^S} \int_0^1 [\sum_{i\in N, a_i\in A_i} (\phi_i(s) + w_i) a_i g_i(a_i|s)]\, p(s)\, ds \tag{OA:6}
$$

This problem can be solved pointwise by strategy $g_i(a_i|s) = 1$ if and only if $a_i \in \arg\max_{\tilde{a}_i \in A_i}(\phi_i(s) + w_i)\tilde{a}_i$. By IVV, $\phi_i(s) + w_i$ is strictly increasing in $s \in S$. Thus either we have $\phi_i(\hat{s}_i) + w_i = 0$ for some $\hat{s}_i \in [0, 1]$, or either $\phi_i(s) + w_i < 0$ for all $s \in S$ or $\phi_i(s) + w_i > 0$ for all $s \in S$. In the first case the maximizing strategy on dimension $i \in N$ is $g_i(a_i|s) = \mathbb{1}\{s \geq \hat{s}_i\}$, where without loss of generality we set $g_i(a_i|\hat{s}_i) = 1$ since $\hat{s}_i \in S$ has measure zero. In the other cases we can write the maximizing strategy as having a threshold form with thresholds $\hat{s}_i = 0$ and $\hat{s}_i = 1$ respectively. □

C Sum-Narrow Participation Constraints

In this section I consider an alternative to narrow dimension-by-dimension participation constraints. The *sum-narrow participation constraint* requires that $\sum_{i \in N} \overline{U}_i(s)$ is greater than the value of the outside option for all types $s \in S$. This is a relaxation of the narrow participation constraint considered earlier. I show how we can modify the analysis of the principal's optimal mechanism for this setting.

The sum-narrow participation constraint holds if for all $s \in S$

$$\sum_{i \in N} \overline{U}_i(s) = \sum_{i \in N} \sum_{a_i \in A_i} g_i(a_i|s)[v_i(s)a_i + \overline{t}_i(a_i)] \geq 0 \qquad \text{(OA:7)}$$

This constraint reflects that the agent makes a joint coordinated decision across dimensions on whether to participate or not. The agent has the option to take the zero action on all dimensions and reject any transfer given by the mechanism. Under the sum-narrow participation constraint, the agent understands that the transfer function may be interactive across dimensions and they cannot just reject the transfer on each dimension individually. This contrasts with their beliefs formed from narrow inference for different actions within the mechanism, which could be correct only if the transfer function was additive. Note that this joint participation constraint is weaker than the narrow participation constraints: if the narrow participation constraints hold on every dimension then so does the joint participation constraint.

I modify the proof of Theorem 1 to obtain the following result.

Theorem 2. *Assume IVV holds. The principal maximizes their objective over all NIC mechanisms that satisfy the sum-narrow participation constraint if and only* *if they choose a transfer function that implements a threshold strategy that solves*

$$\max_{\hat{s}\in[0,1]^n} \overline{W}(\hat{s};\frac{1}{n}) = \max_{\hat{s}\in[0,1]^n}\{\sum_{i\in N}\int_{\hat{s}_i}^{1}(\frac{1}{n}\phi_i(s)+w_i)\,p(s)\,ds\} \tag{OA:8}$$

Proof. The same proof as in Theorem 1 applies up to Step 2. Step 2 still applies but the upper bound problem modified so that we no longer have $\overline{t}(0) \in \mathbb{R}^n_{\geq 0}$, but instead $\overline{t}(0) \in \{t \in \mathbb{R}^n : \sum_{i\in N} t_i = 0\} \equiv \tau^{SN}$. The upper bound problem is then

$$\begin{aligned}
&\min_{\beta\in[0,1]^n,\sum_{i\in N}\beta_i=1}\ \sup_{\hat{s}\in[0,1]^n,\overline{t}(0)\in\tau^{SN}} \overline{W}(\hat{s},\overline{t}(0),\beta)\\
=&\sup_{\hat{s}\in[0,1]^n,\overline{t}(0)\in\tau^{SN}} \overline{W}(\hat{s},\overline{t}(0),\beta^*)\\
=&\sup_{\hat{s}\in[0,1]^n,\overline{t}(0)\in\tau^{SN}}\ \min_{\beta\in[0,1]^n,\sum_{i\in N}\beta_i=1} \overline{W}(\hat{s},\overline{t}(0),\beta)
\end{aligned}$$

with

$$\overline{W}(\hat{s},\overline{t}(0),\beta) = \sum_{j\in N}[-\beta_j\overline{t}_j(0) + \int_{\hat{s}_i}^{1}(\beta_j\phi_i(s)+w_i)\,p(s)\,ds]$$

The saddle point values β^* must be such that $\beta^*_i = \frac{1}{n}$ for all $i \in N$. Otherwise if $\beta^*_j > \frac{1}{n}$ for some $j \in N$ then for any $t < 0$, we can choose $\overline{t}_j(0) = t$, and for all $l \in N\setminus\{j\}$ $\overline{t}_l(0) = -\frac{1}{n-1}t$. This satisfies the sum-narrow participation constraint and allows us to obtain an arbitrarily large payoff by choosing t.

With $\beta^* = (\frac{1}{n},...,\frac{1}{n})$, we can attain the value of the upper bound in the full problem by setting $\overline{t}(0)$ such that $\sum_{i\in N}\overline{t}_i(0) = 0$ and for any $i, j \in N$ the statistical correctness constraint $\overline{t}_i(0)-v_i(s)(1-P(\hat{s}_i)) = \overline{t}_j(0)-v_j(s)(1-P(\hat{s}_j))$ holds. This can be achieved by

$$\overline{t}_i(0) = v_i(s)(1-P(\hat{s}_i)) - \frac{1}{n}\sum_{j\in N} v_j(s)(1-P(\hat{s}_j)) \text{ for all } i \in N$$

□

The only subsequent result that doesn't hold under sum-narrow participation constraints is Proposition 7. Propositions 3, 4, 5 and 6 continue to hold, and their proofs require no significant modification.

D Dropping the IVV assumption

I now explore what happens to the principal's optimal mechanism without the IVV assumption. I first present an example where IVV does not hold, and show the principal's optimal mechanism under the rational benchmark is no longer separable and no longer implements a threshold strategy. Thus, since only threshold strategies are NIC, we cannot make the same neat comparisons between the principal's optimal mechanism under narrow and rational inference that we made in Propositions 3 and 4. Without IVV, the fact that there are strategies that are IC but not NIC becomes relevant.

However, despite the reduction in the set of strategies that are implementable we can prove that an analogous result to Proposition 3 still holds. The principal is still always better off when agents make narrow compared to rational inference if $v_i(s) \leq 0$, $w_i > 0$ for all $i \in N$, $s \in S$ and the principal is restricted to implement a *deterministic interval strategy*. In a deterministic interval strategy, for some $k \in \mathbb{N}$ there is a partition of the type space $z_0 = 0 < z_1 < ... < z_{k-1} < z_k = 1$ where for any $s \in [z_{l-1}, z_l)$ for $l \in \{1, ..., k-1\}$ (and $[z_{k-1}, 1]$ for $l = k$) we have $g(a|s) = 1$ for some $a \in A$.

Finally, I show how the characterization result Theorem 1 can be modified to deal with the absence of IVV.

D.0.1 Example without IVV

Example 3. There are two dimensions $N = \{1, 2\}$ and the type distribution is uniform $P(s) = s$. Let the predictable utility take the following form on either dimension, for some d_i, b_i and $r_i \in (0, 1)$

$$v_i(s) = \begin{cases} (d_i - b_i)(1 - r_i) - \frac{d_i}{2}(1 - s) - \frac{d_i - b_i}{2}\frac{(1-r_i)^2}{1-s} \text{ if } s \in [0, r_i) \\ -\frac{b_i}{2}(1 - s) \text{ if } s \in [r_i, 1] \end{cases} \tag{OA:9}$$

this has a continuous derivative equal to

$$v_i'(s) = \begin{cases} \frac{d_i}{2} - \frac{d_i - b_i}{2}(\frac{1-r_i}{1-s})^2 \text{ if } s \in [0, r_i) \\ \frac{b_i}{2} \text{ if } s \in [r_i, 1] \end{cases} \tag{OA:10}$$

Assume that $d_i(1-(1-r_i)^2)+b_i(1-r_i)^2>0$ and $b_i>0$ so that $v_i(.)$ is strictly increasing. We can write the virtual values associated with these predictable utilities.

$$v_i(s)-\frac{1-P(s)}{p(s)}v_i'(s)=\begin{cases}(d_i-b_i)(1-r_i)-d_i(1-s) \text{ if } s\in[0,r_i)\\ -b_i(1-s) \text{ if } s\in[r_i,1]\end{cases} \tag{OA:11}$$

These virtual values can be decreasing for some interval of types depending on the parameters. Let $d_1=b_1=0.54$, $r_1=0.6$, $d_2=-0.12$, $b_2=1.1$, $r_2=0.66$, $w_1=0.4266$ and $w_2=0.32$. I show that under these parameters the principal's optimal separable mechanism is dominated by a non-separable mechanism implementing a non-threshold strategy.

First calculate the best case mechanism for the principal facing a rational agent if they were restricted to separable mechanisms. For these parameters, the optimal thresholds are interior and solve $\phi_i(\hat{s}_i)+w_i=0$ for $i\in N$.

$$\hat{s}_1^*=1-\frac{w_1}{b_1}=0.21$$
$$\hat{s}_2^*=1-\frac{w_2}{b_2}=\frac{39}{55}$$

The value of the principal's objective under this mechanism is then $W(\hat{s}_1^*,\hat{s}_2^*)\approx 0.215$.

The following mechanism is better for the principal. It implements the deterministic interval strategy g^{int}

$$\begin{aligned}&g^{int}(a_1,a_2|s)\\&=\mathbb{1}\{s\in[0,\hat{s}_1^*)\}(1-a_1)\cdot a_2+\mathbb{1}\{s\in[\hat{s}_1^*,\hat{s}_2^*)\}a_1\cdot(1-a_2)+\mathbb{1}\{s\in[\hat{s}_2^*,1]\}a_1\cdot a_2\end{aligned}$$

This gives the principal payoff

$$\begin{aligned}W(g^{int})&=\int_0^{\hat{s}_1^*}(\phi_2(s)+w_2)\,p(s)ds+\int_{\hat{s}_1^*}^{\hat{s}_2^*}(\phi_1(s)+w_1)\,p(s)ds\\&+\int_{\hat{s}_2^*}^{1}((\phi_1(s)+w_1)+(\phi_2(s)+w_2))\,p(s)ds\\&\approx 0.217\end{aligned}$$

which is greater than the payoff from the best-case for a separable mechanism.

The strategy g^{int} can be made both IC and to satisfy the participation constraint by the following non-separable transfer function.

$$\begin{aligned}
t(0,0) &= 0 \\
t(0,1) &= t(0,0) - v_1(0) \\
t(1,0) &= t(0,1) - v_1(\hat{s}_1^*) \\
t(1,1) &= t(1,0) - v_2(\hat{s}_2^*)
\end{aligned}$$

△

D.0.2 Welfare of the Principal without IVV

I now show that Proposition 3— which distinguishes cases where the principal benefits from facing a narrow agent— also holds without the IVV assumption. In the case where the principal is worse-off under narrow inference, the result is trivial as the principal may now benefit from implementing a larger set of strategies in the rational case compared to the narrow case. The interesting case is the one where the principal gains from narrow inference, where actions have a predictable cost to the agent but a direct benefit to the principal.

The proof for the second case works as follows. For a fixed interval strategy, for each dimension $i \in N$ take the lowest type that takes the action $a_i = 1$. We then take the dimension i^* at which such a threshold has the greatest cost to the principal in terms of the transfer function required for IC. We can obtain an upper bound on the loss of payoff to the principal if they moved to a threshold strategy where only the action on dimension i^* is taken, and only by types above threshold $\underline{s}_{i^*}$. Under narrow inference, for exactly the same cost the principal has to pay to implement this solo action strategy under the rational benchmark, the principal can implement a threshold strategy where substantial payoff from other dimensions is obtained for free. The gain from this move to narrow inference exceeds the upper bound on the loss from moving to the solo action strategy. This shows the principal's payoff from any interval strategy in the rational benchmark is lower than their payoff from some threshold strategy under narrow inference.

Proposition 8. *Assume the principal is restricted to implementing a deterministic interval strategy*

1. *When for every $i \in N$ we have $v_i(s) \leq 0$ for all $s \in S$, the principal can*

obtain at least as high an objective value when the agent is narrow compared to the rational benchmark.

2. *When for every $i \in N$ we have $v_i(s) \geq 0$ for all $s \in S$, the principal obtains at least as high an objective value in the rational benchmark compared to when the agent is narrow.*

Proof. For the second case where $v(s) \geq 0$ for all $s \in S$ and the principal is worse off under narrow inference, from Proposition 3 the result holds if the principal is restricted to implement a threshold strategy. Since the principal can also implement a non-threshold strategy in the rational benchmark, there is then an additional gain to the principal from rational inference over narrow inference without IVV.

For the first case, let g^{int} be an interval strategy with k intervals. Let $N_l \subseteq N$ be the subset of dimensions such that the agent takes the action $a_i = 1$ for some type in interval l; if $i \in N_l$ then $g_i^{int}(a_i|s) = 1$ for all $s \in [z_{l-1}, z_l)$.

For each dimension $i \in N$, we define the smallest type $\underline{s}_i$ that takes the action $a_i = 1$ under g^{int}; $\underline{s}_i = \min\{s \in S : g_i^{int}(1|s) = 1\}$. Each $\underline{s}_i$ is the lower bound of one of the k intervals of g^{int}, for each $i \in N$ denote this interval by $[z_{l_i-1}, z_{l_i})$. The set of dimensions at which action 1 is taken in this interval is denoted N_{l_i} and includes i. Let $i^* = \arg\min_{i \in N} v_i(\underline{s}_i)(1 - P(\underline{s}_i))$.

The principal's welfare when implementing deterministic interval strategy g^{int} in the rational benchmark can be obtained from the expression (OA:6).

$$\sum_{l=1}^{k} \sum_{i \in N_l} [(w_i + v_i(z_{l-1}))(1 - P(z_{l-1})) - (w_i + v_i(z_l))(1 - P(z_l))] \qquad \text{(OA:12)}$$

We can write the principal's payoff from action $a_i = 1$ in interval $[z_{l-1}, z_l)$ as

$$\begin{aligned}&(w_i + v_i(z_{l-1}))(1 - P(z_{l-1})) - (w_i + v_i(z_l))(1 - P(z_l)) \\ &= (w_i + v_i(z_{l-1}))(P(z_l) - P(z_{l-1})) - (v_i(z_l) - v_i(z_{l-1}))(1 - P(z_l))\end{aligned} \qquad \text{(OA:13)}$$

and then rewrite the principal's welfare

$$\sum_{l=1}^{k} \sum_{i \in N_l} (w_i + v_i(z_{l-1}))(P(z_l) - P(z_{l-1})) - \sum_{l=1}^{k} (1 - P(z_l)) \sum_{i \in N_l} (v_i(z_l) - v_i(z_{l-1})) \qquad \text{(OA:14)}$$

We can write the second term in (OA:14) as

$$\begin{aligned}&\sum_{l=1}^{k}(1-P(z_l))\sum_{j\in N_l}(v_j(z_l)-v_j(z_{l-1}))\\&=\sum_{l=1}^{k}(P(z_l)-P(z_{l-1}))\sum_{m=1}^{l}\sum_{j\in N_m}(v_j(z_m)-v_j(z_{m-1}))\end{aligned}\tag{OA:15}$$

For g^{int} to be IC, continuity of $v_i(.)$ and the cyclical monotonicity condition in Lemma B.1 requires that for any intervals $l\in\{1,...,k\}$ and $h\in\{1,...,l\}$

$$\sum_{m=h+1}^{l}\sum_{i\in N_m}(v_i(z_m)-v_i(z_{m-1}))\geq\sum_{m=h}^{l}\sum_{j\in N_h}(v_j(z_m)-v_j(z_{m-1}))\tag{OA:16}$$

$$=\sum_{j\in N_h}(v_j(z_l)-v_j(z_h))\tag{OA:17}$$

In particular, this applies to the first interval that the action $a_{i^*}=1$ is taken; $h=l_{i^*}$. Since for any $h\in\{1,...,k\}$

$$\begin{aligned}&\sum_{l=h}^{k}(P(z_l)-P(z_{l-1}))\sum_{m=h}^{l}\sum_{j\in N_h}(v_j(z_m)-v_j(z_{m-1}))\\&=\sum_{l=h}^{k}(1-P(z_l))\sum_{j\in N_h}(v_j(z_l)-v_j(z_{l-1}))\end{aligned}\tag{OA:18}$$

combining (OA:15), (OA:17), (OA:18) and that from the fact $v_i(.)$ is increasing we have for any $l\in\{1,\ldots,k\}$

$$\sum_{m=1}^{l}\sum_{j\in N_m}(v_j(z_m)-v_j(z_{m-1}))\geq 0$$

then gives us

$$\begin{aligned}&\sum_{l=1}^{k}(1-P(z_l))\sum_{j\in N_l}(v_j(z_l)-v_j(z_{l-1}))\\&\geq\sum_{l=l_{i^*}}^{k}(1-P(z_l))\sum_{i\in N_{l_{i^*}}}(v_i(z_l)-v_i(z_{l-1}))\end{aligned}$$

and thus from (OA:14) the following upper bound on the welfare of the principal.

$$\sum_{l=1}^{k}\sum_{i\in N_l}(w_i+v_i(z_{l-1}))(P(z_l)-P(z_{l-1}))-\sum_{l=l_{i^*}}^{k}(1-P(z_l))\sum_{i\in N_{l_{i^*}}}(v_i(z_l)-v_i(z_{l-1})) \tag{OA:19}$$

Since $v_i(.)$ is increasing and $v_i(s)\leq 0$ for all $s\in S$, this in turn is upper bounded by

$$\begin{aligned}
&\sum_{l=1}^{k}\sum_{i\in N_l}(w_i+v_i(z_{l-1}))(P(z_l)-P(z_{l-1}))-\sum_{l=l_{i^*}}^{k}(v_{i^*}(z_l)-v_{i^*}(z_{l-1}))(1-P(z_l))\\
&\leq \sum_{j\in N} w_j(1-P(\underline{s}_j)) && \text{(OA:20)}\\
&+\sum_{l=l_{i^*}}^{k}[v_{i^*}(z_{l-1})(P(z_l)-P(z_{l-1}))-(v_{i^*}(z_l)-v_{i^*}(z_{l-1}))(1-P(z_l))] && \text{(OA:21)}\\
&=\sum_{j\in N} w_j(1-P(\underline{s}_j))+\int_{\underline{s}_{i^*}}^{1}(v_{i^*}(s)-\frac{1-P(s)}{p(s)}v_{i^*}'(s))\,p(s)ds && \text{(OA:22)}
\end{aligned}$$

Under narrow inference, the principal can achieve this upper bound by implementing a threshold strategy g^* such that $g_i^*(1|s)=\mathbb{1}\{s\geq \underline{s}_i\}$ for all $i\in N$. This gives the result.

The principal does this by implementing beliefs such that $\overline{t}_{i^*}(0)=0$, $\overline{t}_{i^*}(1)=-v_{i^*}(\underline{s}_{i^*})$, and for all $j\in N\setminus\{i^*\}$

$$\begin{aligned}
\overline{t}_j(0)&=v_j(\underline{s}_j)(1-P(\underline{s}_j))-v_{i^*}(\underline{s}_{i^*})(1-P(\underline{s}_{i^*}))\geq 0\\
\overline{t}_j(1)&=\overline{t}_j(0)-v_j(\underline{s}_j)
\end{aligned}$$

from Proposition 2 we can find a transfer function that implements these beliefs since g^* is a threshold strategy. □

D.1 Minimax characterization without IVV

In this section I show how we can adapt Theorem 1 when IVV does not necessarily hold. This works by applying the ironing procedure of Myerson (1981). Let $P^{-1}(s)$ be the quantile function for the type distribution.[1] For any dimension $i\in N$ and

[1] This is the inverse of cdf P as P is strictly increasing, and thus P^{-1} is also strictly increasing.

any $x \in [0,1]$, define

$$\Phi_i(x) = \int_x^1 \phi_i(P^{-1}(u))\,du \tag{OA:23}$$

Let $co(\Phi_i)$ be convex hull of the graph of Φ_i, the upper concave envelope of $\Phi_i(.)$ is then defined as

$$\widehat{\overline{\Phi}}_i(x) = \sup\{z : (z,x) \in co(\Phi_i)\} \tag{OA:24}$$

Then let $\widehat{\overline{\phi}}_i(s) = -\widehat{\overline{\Phi}}'_i(P(s))$ and $\widehat{\overline{\Phi}}_i(x) = \int_x^1 \widehat{\overline{\phi}}_i(P^{-1}(x))$, this is well defined and increasing for all $s \in S$ and $x \in [0,1]$ by definition of $\widehat{\overline{\Phi}}_i(.)$.[2] For any $\beta \in [0,1]^n$ and dimension $i \in N$, define the threshold[3]

$$\hat{s}_i^*(\beta) = \min\{\tilde{s} \in \underset{\hat{s}_i \in S}{\arg\max} \int_{\hat{s}_i}^1 (\beta_i \widehat{\overline{\phi}}_i(s) + w_i)\, p(s)\,ds\}$$

and let $\hat{s}^*(\beta) = (\hat{s}_i^*(\beta))_{i \in N}$. We can now state the result.

Theorem 3. *The principal maximizes their objective over all NIC mechanisms that satisfy the narrow participation constraints if and only if they choose a transfer function that implements a threshold strategy $\hat{s}^*(\beta^*)$, where*

$$\beta^* \in \underset{\beta \in [0,1]^n : \sum_{i \in N} \beta_i = 1}{\arg\min} \sum_{i \in N} \int_{\hat{s}_i^*(\beta)}^1 (\beta_i \widehat{\overline{\phi}}_i(s) + w_i)\, p(s)\,ds \tag{OA:25}$$

and the value of the principal's objective is given by

$$\widehat{\overline{W}}(\hat{s}^*(\beta^*); \beta^*) = \sum_{i \in N} \int_{\hat{s}_i^*(\beta^*)}^1 (\beta_i^* \widehat{\overline{\phi}}_i(s) + w_i)\, p(s)\,ds \tag{OA:26}$$

Proof. Step 1 of Theorem 1 works as before. From Step 2 onwards, we replace the objective by

$$\widehat{\overline{W}}(x, \overline{t}(0), \beta) = \sum_{j \in N} [-\beta_j \overline{t}_j(0) + \int_{x_j}^1 (\beta_j \widehat{\overline{\phi}}_j (P^{-1}(u)) + w_j)du]$$

where $P^{-1}(.)$ is the strictly increasing quantile function for the type distribution

[2] We have that by concavity $\widehat{\overline{\Phi}}'_i(P(.))$ is defined for all but a countable set of points in S, and by right continuity we can extend it to all S.

[3] This is well defined due to continuity of $\int_{\hat{s}_i}^1 (\beta_i \widehat{\overline{\phi}}_i(s) + w_i)\, p(s)\,ds$.

as before. This is an upper bound on the original objective as by definition of the upper concave envelope, $\int_x^1(\hat{\hat{\phi}}_i(P^{-1}(u))du = \hat{\hat{\Phi}}_i(x) \geq \Phi_i(x) = \int_x^1(\phi_i(P^{-1}(u))du$ for all $i \in N$, $x \in [0,1]$. Since by Lemma 2 only threshold strategies are NIC and all threshold strategies can be made NIC by some transfer function, the new objective remains an upper bound of the full problem even without increasing $\phi_i(.)$.

Since $\hat{\hat{\phi}}_i(.)$ is increasing in s the new objective is concave for fixed β, and we can make the same argument as we made for Theorem 1 using the minimax theorem to get

$$\begin{aligned}
&\min_{\beta\in[0,1]^n, \sum_{i\in N}\beta_i=1} \sup_{\hat{s}\in[0,1]^n, \bar{t}(0)\in\mathbb{R}^n_{\geq 0}} \widehat{\widehat{W}}(\hat{s}, \bar{t}(0), \beta) \\
&= \sup_{\hat{s}\in[0,1]^n, \bar{t}(0)\in\mathbb{R}^n_{\geq 0}} \widehat{\widehat{W}}(\hat{s}, \bar{t}(0), \beta^*) \\
&= \sup_{\hat{s}\in[0,1]^n, \bar{t}(0)\in\mathbb{R}^n_{\geq 0}} \min_{\beta\in[0,1]^n, \sum_{i\in N}\beta_i=1} \widehat{\widehat{W}}(\hat{s}, \bar{t}(0), \beta)
\end{aligned}$$

In $\widehat{\widehat{W}}(\hat{s}, \bar{t}(0), \beta)$ the term for $\bar{t}(0)$ is separable, therefore $\hat{s}^*(\beta) \in \arg\max_{\hat{s}\in[0,1]^n} \widehat{\widehat{W}}(\hat{s}, \bar{t}(0), \beta)$ for all $\bar{t}(0)$. Thus, these thresholds maximize the new objective for fixed β. Therefore, we must have $\sup_{\hat{s}\in[0,1]^n, \bar{t}(0)\in\mathbb{R}^n_{\geq 0}} \widehat{\widehat{W}}(\hat{s}, \bar{t}(0), \beta^*) = \sup_{\bar{t}(0)\in\mathbb{R}^n_{\geq 0}} \widehat{\widehat{W}}(\hat{s}^*(\beta^*), \bar{t}(0), \beta^*)$.

We now apply the same argument as Myerson (1981). For any $i \in N$

$$\begin{aligned}
&\int_{\hat{s}_i}^1 (\phi_i(s) - \hat{\hat{\phi}}_i(s))\, p(s) ds = \int_{P(\hat{s}_i)}^1 (\phi_i(P^{-1}(u)) - \hat{\hat{\phi}}_i(P^{-1}(u)))du \\
&= \Phi_i(P(\hat{s}_i)) - \hat{\hat{\Phi}}_i(P(\hat{s}_i))
\end{aligned}$$

We can show that $\Phi_i(P(\hat{s}_i^*(\beta))) = \hat{\hat{\Phi}}_i(P(\hat{s}_i^*(\beta)))$ for any β. When $\hat{s}_i^*(\beta) \in \{0,1\}$ this is clear. If $\hat{s}_i^*(\beta) \in (0,1)$ then $\beta_i\hat{\hat{\phi}}_i(\hat{s}_i^*(\beta)) + w_i = 0$, and by definition $\hat{s}_i^*(\beta)$ is the smallest type satisfying this. Since $\hat{\hat{\Phi}}_i(P(s)) > \Phi_i(P(s))$ only in intervals $s \in [\underline{s}, \overline{s})$ where $\hat{\hat{\phi}}_i(s)$ is constant, at $\hat{s}_i^*(\beta)$ we must have $\Phi_i(P(\hat{s}_i^*(\beta))) = \hat{\hat{\Phi}}_i(P(\hat{s}_i^*(\beta)))$ otherwise we can find a smaller threshold in the maximizing set.

For any β we can write the old objective as

$$
\begin{aligned}
&\overline{W}(\hat{s}, \overline{t}(0), \beta) \\
&= \sum_{j \in N} [-\beta_j \overline{t}_j(0) + \int_{\hat{s}_j}^{1} (\beta_j \phi_j(s) + w_j) p(s) ds] \\
&= \sum_{j \in N} [-\beta_j \overline{t}_j(0) + \int_{\hat{s}_j}^{1} (\beta_j \widehat{\overline{\phi}}_j(s) + w_j) p(s) ds \\
&+ \beta_j \int_{\hat{s}_j}^{1} (\phi_j(s) - \widehat{\overline{\phi}}_j(s)) p(s) ds] \\
&= \sum_{j \in N} [-\beta_j \overline{t}_j(0) + \int_{\hat{s}_j}^{1} (\beta_j \widehat{\overline{\phi}}_j(s) + w_j) p(s) ds \\
&+ \beta_j (\Phi_j(P(\hat{s}_j)) - \widehat{\overline{\Phi}}_j(P(\hat{s}_j)))]
\end{aligned}
$$

At $\hat{s}^*(\beta)$, since $\Phi_i(P(\hat{s}_i^*(\beta))) = \widehat{\overline{\Phi}}_i(P(\hat{s}_i^*(\beta)))$ the value the old and new objectives are identical for any β; $\overline{W}(\hat{s}^*(\beta), \overline{t}(0), \beta) = \widehat{\overline{W}}(\hat{s}^*(\beta), \overline{t}(0), \beta)$. This means that for any $\hat{s} \in [0, 1]$

$$
\overline{W}(\hat{s}^*(\beta), \overline{t}(0), \beta) = \widehat{\overline{W}}(\hat{s}^*(\beta), \overline{t}(0), \beta) \geq \widehat{\overline{W}}(\hat{s}, \overline{t}(0), \beta) \geq \overline{W}(\hat{s}, \overline{t}(0), \beta)
$$

which then implies

$$
\begin{aligned}
& \sup_{\hat{s} \in [0,1]^n, \overline{t}(0) \in \mathbb{R}^n_{\geq 0}} \; \min_{\beta \in [0,1]^n, \sum_{i \in N} \beta_i = 1} \overline{W}(\hat{s}, \overline{t}(0), \beta) \\
&= \sup_{\overline{t}(0) \in \mathbb{R}^n_{\geq 0}} \; \min_{\beta \in [0,1]^n, \sum_{i \in N} \beta_i = 1} \overline{W}(\hat{s}^*(\beta^*), \overline{t}(0), \beta) \\
&= \sup_{\overline{t}(0) \in \mathbb{R}^n_{\geq 0}} \; \min_{\beta \in [0,1]^n : \sum_{i \in N} \beta_i = 1} \widehat{\overline{W}}(\hat{s}^*(\beta^*), \overline{t}(0), \beta) \\
&= \min_{\beta \in [0,1]^n, \sum_{i \in N} \beta_i = 1} \; \sup_{\hat{s} \in [0,1]^n, \overline{t}(0) \in \mathbb{R}^n_{\geq 0}} \widehat{\overline{W}}(\hat{s}, \overline{t}(0), \beta) \\
&= \sup_{\hat{s} \in [0,1]^n, \overline{t}(0) \in \mathbb{R}^n_{\geq 0}} \widehat{\overline{W}}(\hat{s}, \overline{t}(0), \beta^*) \\
&= \sup_{\overline{t}(0) \in \mathbb{R}^n_{\geq 0}} \overline{W}(\hat{s}^*(\beta), \overline{t}(0), \beta^*) \\
&= \min_{\beta \in [0,1]^n : \sum_{i \in N} \beta_i = 1} \; \sup_{\hat{s} \in [0,1]^n, \overline{t}(0) \in \mathbb{R}^n_{\geq 0}} \overline{W}(\hat{s}, \overline{t}(0), \beta)
\end{aligned}
$$

giving us a minimax result for $\overline{W}(.)$ also. We can then apply Step 3 from Theorem

1 to show that this upper-bound objective value can be achieved by a particular transfer function and the beliefs it induces, which completes the proof. □