Leveraging Benchmarks via Information Design[*]

Christopher Teh[†]

This version: August 26, 2024 (Latest)

Abstract

An agent's reasoning about how others behave is often benchmarked on payoff-irrelevant factors such as culture, biases and past performance. I study how these benchmarks shape a designer's optimal information disclosure in symmetric binary action supermodular games. Agents determine their (Bayes-Nash) equilibrium actions via introspection, anchored on a benchmark of aggregate behaviour. I characterize all outcomes implementable by some information structure, and use the result to solve the designer's problem. I show that the designer benefits from a higher benchmark about aggregate behaviour, and she leverages it by disclosing information more symmetrically among agents. Meanwhile, higher benchmarks may not benefit agents. I also find that changes in the benchmark have a greater impact when the designer limited to public information. My approach nests prior analyses using designer-preferred or adversarial selection, which coincides with extreme benchmarks.

Keywords: Information Design, Supermodular Games, Equilibrium Selection

[*] I thank my advisors Richard Holden, Keiichi Kawai and Anton Kolotilin for their continual guidance and support, and participants at the SAET 2023, ESAM Sydney 2023, AETW 2024, and local workshops at UNSW Sydney and the Toulouse School of Economics (TSE) for providing insightful comments. This paper has also benefited greatly from discussions with many local Academics and PhD students during my visit to TSE in the fall of 2023, of which I am indebted to. Online Supplementary Appendix can be found here

[†] UNSW Sydney, chris.teh@unsw.edu.au

1 Introduction

In many economic settings, a designer discloses information to multiple agents engaging in a coordination game. For instance, a policy maker designs stress tests which reveal information about a bank's fundamentals to depositors to dissuade them from making a run on the bank. An entrepreneur decides what information to disclose about her project in her pitch to potential investors to raise funds. A manager decides what information to reveal to employees working on complementary tasks, e.g., about task difficulty and rewards, to motivate them to work.

Because agents' actions are strategic complements, whether agents' coordinate on a given action depend not only on the information received, but also on what they reason others will do. In light of strategic uncertainty, agents often base their reasoning on a *benchmark* about others' behaviours. These benchmarks are shaped by a variety of payoff irrelevant factors. For example, a depositor's benchmark about the number of other depositors making a run on the bank depends on the bank's reputation or historical data about depositor behaviour. An investor's reasoning about whether the entrepreneur can raise enough funds from other investors to launch the project can be shaped by perceived biases given the entrepreneur's attractiveness, race and gender.[1] A worker's reasoning about whether others exert effort depends on the organization's culture surrounding work. Thus, different benchmarks can lead to agents reacting very differently, even to the same piece of information.[2]

The above leads to several natural questions. What information should the designer disclose for a given benchmark? How does it vary in the benchmark? What consequences does this have on outcomes? In this paper, I address these questions in a general *binary action supermodular* setting, which models all of the examples above. I present a systematic approach for incorporating benchmarks into the information design problem, and study how they shape optimal information disclosure.

In the model, a designer first commits to an information structure, which reveals information about an unknown state-of-the-world to a unit mass of symmetric agents.

[1] As noted by Addy Miise, co-Founder and CEO of Jetstream Africa: *"In my experience..., I have found that many...mainstream start-up investors approach investing like judges in the beauty pageant problem. Even when they see inherent potential in great, underrepresented founders, they discount that potential because they don't think the other judges will see it. They believe that the biases of the past will control outcomes in the future."* (Miishe, 2022)

[2] For example, Brooks et al. (2014) find that pitches by attractive, male entrepreneurs to be more persuasive than those by other counterparts, even when the content was the same.

Each agent then simultaneously chooses between one of two actions: invest or not invest. The designer always prefers higher over lower aggregate investment. Likewise, an agent's incentive to invest over not is increasing in the aggregate investment.

The designer optimizes over information structures with a *Monotone Introspective Equilibrium* (MIE) (Kets and Sandroni, 2020, Akerlof and Holden, 2019), a type of Bayes Nash equilibrium, given the *benchmark*. One first computes a sequence of best-responses for agents: a "round one (of introspection)" best-response to a common belief about aggregate investment on each state called the benchmark, a "round two" best-response to other agents following the round one best-response, and so on. If agents only switch from not investing to investing during the sequence, then call the limit strategy the MIE, which is how agents behave under the information structure. Notably, the designer's problem under MIE for extreme benchmarks coincides with that under designer-preferred and adversarial equilibrium selection, the focus of the literature (Bergemann and Morris, 2019). Thus, the problem I study nests prior work.

My main methodological contribution is to provide the means for solving the designer's problem under MIE. My approach relies on the observation that agents who invest in a MIE fall into one of two roles, depending on "when" the agent switches to invest during introspection. First, there are *anchors*: agents who invest against the benchmark and so immediately switch to invest. Second, there are *non-anchors*, agents who do not invest against the benchmark, but switch eventually after enough rounds of introspection. Every information structure implements a distribution over roles across states, called an *introspective outcome*. Thus, the designer's problem can be reformulated as optimizing over implementable introspective outcomes.

To use the reformulation, I characterize all implementable introspective outcomes. I show that these are pinned down by introspective *obedience constraints*, which reflect the incentives of agents to invest in each role. I also offer a class of information structures which support all implementable introspective outcomes. Under it, anchors are provided symmetric information, while different non-anchors are provided different information. This gives a canonical interpretation for the choice of an introspective outcome: it captures how much symmetric (anchors) and asymmetric (non-anchors) information the designer provides to induce investment across states.

I use my approach to solve the information design problem in *threshold games*. These are games where an agent's payoff from investing is constant in the aggregate investment whenever it is negative. These span many settings, for instance, team production models and regime change games (Morris and Shin, 2003). For these, I explicitly construct an optimal introspective outcome. I show that under it, a higher benchmark raises the aggregate investment on all states, and the mass of anchors on most states. The former implies the designer benefits from a higher benchmark, and she leverages it to alleviate the coordination problem. The latter implies the designer uses more symmetric information to induce investment under a higher benchmark.

I deliver two additional insights about the interaction between benchmarks and the optimal introspective outcome constructed. First, I identify a weak sufficient condition for it to perfectly coordinate investments across all states. It requires that on all states where agents' payoffs from investing given the benchmark are negative, so agents cannot be "easily" induced to invest, the designer's payoff is convex *relative* to that of agents', so she prefers randomizing over having all agents invest and not invest over having a fraction invest. A higher benchmark makes this condition easier to satisfy, and so raises designer's incentive to perfectly coordinate investments. I also show that in many relevant settings, the condition is satisfied under all benchmarks. If so, then changes in the benchmarks only affects the designer's incentive to discriminate through information provision, and not ex-post.

Second, I shed light on whether raising the benchmark increases agents' ex-ante payoffs under the optimal introspective outcome. The key insight is that it depends on which states the benchmark increases on. For example, in the context of an entrepreneur (designer) raising investment from investors (agents), the type of project (state) the entrepreneur is perceived to be better at raising investment for (higher benchmark) dictates whether investors are better or worse off. I characterize the kinds of increases in the benchmark that increase or decrease agents' payoffs. In particular, these results show that one can often find sequences of increases in the benchmark that lead to non-monotone changes in agents' payoffs.

Finally, I consider the implications of limiting the designer to public information. This applies, for instance, to the disclosure of banking stress test results or project information in an IPO. I show that for generic benchmarks, the designer is strictly worse off when restricted to public information than when she is not. Meanwhile, the marginal gain to the designer from raising the benchmark is larger when restricted to public information. This suggests payoff-irrelevant factors have a greater impact on (optimal) outcomes when there are informational restrictions in play.

Related literature. My paper contributes to the vast literature on optimal information design (Bergemann and Morris, 2019, Taneva, 2019), particularly in binary action supermodular (BAS) games. Most if not all papers impose either designer-preferred selection (e.g., Arieli and Babichenko (2019), Candogan and Drakopoulos (2020) and Taneva and Mathevet (2023)), or adversarial selection (e.g., Goldstein and Huang (2016), Mathevet et al. (2020), Li et al. (2023), Hoshino (2022), Inostroza and Pavan (2023) and Morris et al. (2022a, 2024)),[3] and focus on how changes in the base game or informational constraints shape optimal information disclosure.[4] Relative to these, my paper is, to my knowledge, the first to (also) consider "intermediate" selection via intermediate benchmarks. I also develop novel results about comparative statics of optimal information structures in the benchmark, how it affects outcomes, and the interaction between benchmarks and informational restrictions.

My work also speaks to the literature on contracting with externalities (Segal, 1999, Segal, 2003) studying optimal mechanisms for implementing perfect coordination under adversarial equilibrium selection.[5] A key insight is that the designer optimally leverages the iterative dominance reasoning of agents which pins down the adversarial equilibrium by discriminating between agents. For example, by offering fully discriminatory incentives when incentives must be public (Winter, 2004), or offering fully discriminatory information when incentives can be private (Halac et al., 2021, Halac et al., 2022, Morris et al., 2022b) or must be symmetric (Moriya and Yamashita, 2020). Complementary to these, I show that higher benchmarks, i.e., less adversarial selection, lead to more symmetric information provision. In particular, for generic benchmarks, *partial* discrmination is strictly optimal. I also find that higher benchmarks also raise the designer's incentive to treat agents symmetrically ex-post.

I also contribute to the recent literature on introspective equilibrium (Kets and Sandroni, 2020). It offers a payoff-irrelevant alternative for equilibrium selection to

[3] Li et al. (2023), Hoshino (2022) and Morris et al. (2022a) also consider optimal information design when restricted to a finite signal space. The exercise is reminiscent of information design under Level-K thinking (Crawford et al., 2013) for the worst "L0" belief, where equilibrium behaviours are taken as the Kth best response. Meanwhile, introspective equilibrium takes the limiting behaviour as $K \to \infty$, and I study the implications for optimal disclosure varying L0 beliefs (benchmarks).

[4] Relatedly, Lipnowski et al. (2024) study when the designer's equilibrium payoff (and certain features of equilibrium information structures) is robust to varying selection in Bayesian Persuasion. Carroll (2016) and Ziegler (2020) study adversarial information design in bilateral settings.

[5] Traditionally, these papers study the problem when the designer chooses among mechanisms with a unique equilibrium. When the designer's objective is to achieve perfect coordination, this problem coincides with the one under adversarial selection.

the global games approach (Carlsson and van Damme, 1993). Ongoing work applies this to study complete information environments, e.g., capital assembly for projects (Akerlof and Holden, 2019),[6] competition in markets with network externalities (Akerlof et al., 2023), supply chain coordination (Akerlof and Holden, 2023) and organizational design (Kets, 2021). Meanwhile, I extend the introspective equilibrium concept to incomplete information settings where agents share common uncertainty about a payoff-relevant state-of-the-world.[7], and study optimal information design. The main innovation, allowing the benchmark to vary across states, turns out to be crucial for studying how changes in the benchmark affect players' welfare.

My paper also relates to the literature on implementation via information design in games. Aumann (1974) and Bergemann and Morris (2016) characterize the equilibrium outcomes that arise under some information structure in games without and with common state uncertainty respectively. Oyama and Takahashi (2020) and Morris and Ui (2005) provide necessary and sufficient conditions, respectively, for full implementation in complete information BAS games.[8] Morris et al. (2024) extends the construction of Oyama and Takahashi (2020) to fully characterize smallest and unique implementation in incomplete information BAS games. I build on these results by characterizating implementation under intermediate adversariality (benchmarks).

Finally, my work sheds light on how payoff-irrelevant factors such as culture, biases and reputation can shape outcomes. This speaks to broader questions about persistent performance differences among seemingly similar enterprises (Chassang, 2010, Gibbons and Henderson, 2013), how culture affects institutional design (Alesina and Giuliano, 2015) and communication in organizations (Crémer, 1993), and how statistical discrimination (Phelps, 1972, Arrow, 1973) shapes entrepreneurial outcomes.

Organization. Section 2 solves an example to highlight the main insights of the paper. Section 3 introduces the general model. Section 4 characterizes implementable outcomes. Section 5 studies the general information design problem. Section 6 examines the implications of restricting the designer to public information. Section 7 concludes. Formal proofs and additional details are relegated to the Appendix A-C

[6] Akerlof and Holden (2016) use the global games approach to study a similar problem.

[7] In contrast, Kets et al. (2022) only allows for payoff-irrelevant state uncertainty.

[8] Also related are Kajii and Morris (1997) and Rubinstein (1989), who show that risk-dominance is necessary and sufficient for an equilibrium to be fully implementable, i.e., *robust to incomplete information*, in 2×2 complete information games, and Frankel et al. (2003), who provides sufficient conditions for full implementation in general complete information supermodular games.

and the Online Supplementary Appendix.

2 Main Example

An entrepreneur is raising investment for her project from a unit mass of investors. An investor who does not invest receives zero. An investor who chooses to invest incurs a cost of 1, and obtains a yield of 2 conditional on project success. The probability of success depends on both the aggregate investment $I \in [0, 1]$, and the unobserved project quality $\theta \in \{H(\text{igh}), L(\text{ow})\}$. If the quality is high, then the project succeeds with probability $p(I, H) = (3+I)/4$. If the quality is low, then the project always fails: $p(I, L) = 0$. Assume that qualities are equally likely: $Pr(\theta = H) = Pr(\theta = L) = 1/2$.

The entrepreneur decides what verifiable information about θ to gather and disclose in her pitch to investors. I model this as the entrepreneur committing to a pitch structure $\mathcal{S} \equiv (S, \pi(\cdot|H), \pi(\cdot|L))$, delivering pitches $s \in S$ to investors such that the empirical distribution over pitches is $\mu \in \Delta(S)$ with probability $\pi(\mu|\theta)$ on quality θ. The delivery of pitches does not discriminate between investors ex-ante, so each investor privately observes pitch s with probability $\mu(s)$. Finally, investors form beliefs over what other investors observe and θ, and decide whether to invest.[9]

Investors' (Bayes-Nash equilibrium) behaviour under a pitch structure $\mathcal{S}$ is determined by introspection, anchored on a belief about aggregate investment across states $b \in [0, 1]$ called the *benchmark*. More precisely, consider a sequence of "introspective" best-responses defined as follows. The first term in the sequence is the investor's best-response against the benchmark b. Under it, the investor invests if and only if $\mathbb{E}[2p(b, \theta) - 1|s] \geq 0$, where $\mathbb{E}[\cdot|s]$ is the expectation given the investor's belief about (μ, θ) observing s. The $k \geq 2$th term is the investor's best response after k rounds of introspection. Under it, the investor invests if and only if $\mathbb{E}[2p(I^{k-1}(\mu), \theta) - 1|s] \geq 0$, where $I^{k-1}(\mu)$ is the aggregate investment when the distribution over (other) investors' signals is μ, and other investors follow their $k - 1$th round best-response.

[9] Alternatively, the entrepreneur can raise investments on a crowdfunding platform. Here, investors incur a sunk cost from investment when a *flexible funding model* is used (e.g., on Indiegogo), or when opportunity costs are large (e.g., for equity crowdfunding). The simultaneous investment timing captures the idea that most investments are made early or late in the funding cycle (Colombo et al., 2015, Crosetto and Regner, 2018). Assume the platform earns commission on funds raised, so its interests are aligned with the entrepreneur. Then, the platform's design problem, choosing what verifiable information the entrepreneur provides via its information policy, and what projects to privately recommend to investors via its recommender system, mirrors the entrepreneur's.

If an investor only ever switches to investing and never back along the sequence, then call the limit the *b-Monotone Introspective Equilibrium* (b-MIE) of $\mathcal{S}$, which is how investors behave under $\mathcal{S}$.[10] The entrepreneur chooses among pitch structures to maximize the (expected) aggregate investment in her project.

The benchmark b captures investors' reasoning about the entrepreneur's ability to raise investment. This is shaped by exogenous, payoff-irrelevant factors. For instance, investor biases about the entrepreneur's attractiveness, race and gender not correlated with project quality. It can also reflect investors benchmarking on the investment raised by (other) entrepreneurs with similar characteristics in the past. Finally, it can depend on the entrepreneur's or backing venture capitalist reputation.

If the benchmark b is high enough, then disclosing no information induces all investors to invest. To see this, let $b = 1$. The expected payoff from investing against the benchmark is $\mathbb{E}[2p(1, \theta) - 1] = 0$, so an investor invests at the first round of introspection. Anticipating all other investors to do so the same, an investor then invests at the second round of introspection. Repeating the argument, investors invest at all rounds of introspection, and so in the limit, i.e., the b-MIE.

Now consider $b = 1/2$. Here, investors no longer invest under no disclosure, as they are insufficiently optimistic about the quality being high to invest against the benchmark. Thus, the entrepreneur must disclose more information about the project quality to induce investment. More strikingly, the entrepreneur now also finds it optimal to deliver *different* pitches to different investors.

$$\begin{array}{llll} Pr\left(\text{ all observe } s = 1 \ |H\right) & = 1 \qquad & Pr\left(\text{ all observe } s = 1 \ |L\right) & = \dfrac{3}{4} \\ & & Pr\left(\text{ all observe } s = \infty \ |L\right) & = \dfrac{1}{4} \end{array}$$

Figure 1: A symmetric pitch structure for benchmark $b = 1/2$

To see this, suppose the entrepreneur delivers the same pitch to all investors as in Figure 1. Here, all investors are optimistic enough about the quality being high observing pitch $s = 1$ to invest under it, but not under pitch $s = \infty$. More importantly, an investor observing $s = 1$, knowing all other investors invest at all

[10] If one does not exist for b, then investors' behaviours coincide with the largest $\tilde{b}$-MIE satisfying $\tilde{b} \leq b$. As Section 3.3 discusses, the designer always finds a pitch structure with a b-MIE optimal.

rounds of introspection, has a payoff of $\mathbb{E}[2p(1,\theta)-1|1]=1/8>0$ from investing at rounds two and above of introspection, and so strictly prefers to invest then.

$$
\begin{array}{llll}
Pr\left(\begin{array}{c}4/5 \text{ observe } s=1\\ 1/5 \text{ observe } s=2\end{array} |H\right) & =1 & Pr\left(\begin{array}{c}4/5 \text{ observe } s=1\\ 1/5 \text{ observe } s=2\end{array} |L\right) & =\dfrac{3}{4}\\
Pr\left(\text{ all observe } s=2 \ |L\right) & =\dfrac{3}{100} & Pr\left(\text{ all observe } s=\infty \ |L\right) & =\dfrac{22}{100}
\end{array}
$$

Figure 2: A strict improvement over Figure 1 for benchmark $b=1/2$

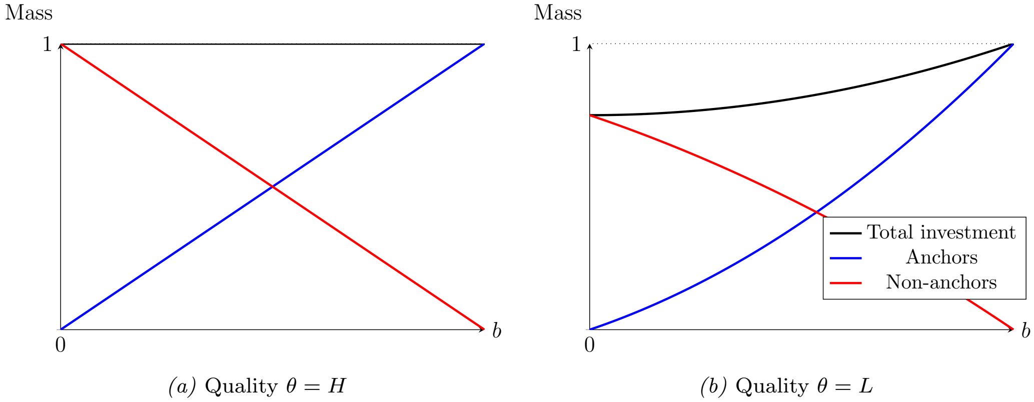
To capitalize on this “slack”, consider the pitch structure in Figure 2. By only slightly reducing the fraction of investors who observe $s=1$ but preserving the frequencies on which it is drawn across qualities, an investor continues to invest at all rounds of introspection under $s=1$. Meanwhile, under the new pitch $s=2$ created, an investor is too pessimistic about project quality to invest against the benchmark. However, because such an investor believes many other investors have observed pitch $s=1$ and so invest at all rounds of introspection, the investor's payoff from investing at rounds two and above of introspection is positive. Thus, the investor switches to investing at the second round of introspection. In turn, all investors invest under pitches $s=1$ and $s=2$. This strictly raises investment in the low quality project.

$$
\begin{array}{llll}
Pr\left(\begin{array}{c}b \text{ anchors, and}\\ 1-b \text{ non-anchors}\end{array} |H\right) & =1 & Pr\left(\begin{array}{c}\frac{2b(1+b)}{3+b^2} \text{ anchors, and}\\ \frac{3-b^2-2b}{3+b^2} \text{ non-anchors}\end{array} |L\right) & =\dfrac{3+b^2}{4}\\
 & & Pr\left(\text{ all investors do not invest } |L\right) & =\dfrac{1-b^2}{4}
\end{array}
$$

Figure 3: Optimal distribution over anchors and non-anchors given benchmark b

The key insight is that the entrepreneur can “use” the investment of *anchors*, investors who prefer to invest against the benchmark (i.e., those who observe pitch $s=1$), to induce *non-anchors* to invest after further introspection (i.e., those who observe pitch $s=2$). The “cost” of using anchors, however, is that anchors require a more optimistic pitch about project quality to invest, which the entrepreneur has limited ability to generate. An optimal pitch structure balances the two.

Using the methods developed in this paper, I derive the distribution over anchors and non-anchors induced by an optimal pitch structure, given in Figure 3. Observe that investors' investments are always *perfectly coordinated*: either all investors invest or all do not. This has to do with the convexity of the entrepreneur's objective: when she cannot easily induce investment in the (low-quality) project, she at least prefers to induce a randomization over all investors investing and not, over having a fraction invest. Hence, investment outcomes remain extreme regardless of the benchmark.



(a) Quality $\theta = H$

(b) Quality $\theta = L$

Figure 4: Investment composition under the optimal distribution over anchors and non-anchors

Meanwhile, as seen in Figure 4, a higher benchmark does raise the probability the low quality project is (fully) funded. Intuitively, a higher benchmark raises anchors' incentives to invest in the high quality project. The designer can leverage this by more frequently pitching a low quality project to investors while maintaining their incentives to invest. This benefits the entrepreneur but harms investors.[11]

Finally, the mass of anchors induced is increasing in the benchmark. Intuitively, by raising investors' payoffs from investing against the benchmark, a higher benchmark also reduces the *relative* "cost" of inducing investment from anchors over non-anchors. This means replacing some non-anchors with anchors is now optimal. As I show later, the optimal distribution can always be implemented by delivering symmetric pitches to anchors, while different non-anchors receive different pitches. Thus, the result states that the entrepreneur now pitches more symmetrically.[12]

[11] This is consistent with some findings from the literature. Ewens and Townsend (2020) find that male investors, who make up a majority of investors in their dataset, express more interest in funding male over female entrepreneur led start-ups, and these biases are not based on inherent gender differences. They also find that female led start-ups, i.e., the ones with a lower benchmark, *outperform* their male counterparts on average. Hebert (2023) finds a similar result.

[12] In the crowdfunding interpretation (see footnote 9), this suggests that projects helmed by entrepreneurs with high benchmarks, e.g, "superstars", should be more widely broadcast (more anchors) and rely less on targeted recommendations (less non-anchors) then those which are not.

3 General Model

3.1 Set-up

A *designer* discloses information about an unknown state of the world $\theta \in [0,1]$ to a unit mass of symmetric *agents* $i \in [0,1]$. Each agent chooses either to invest ($a_i = 1$), or not to invest ($a_i = 0$). An agent's payoff from not investing is zero. Meanwhile, an agent's payoff from investing $u(I,\theta)$, and the designer's payoff $v(I,\theta)$, depend on the aggregate investment $I = \int_i a_i di$ and the state θ. Both $v(I,\theta)$ and $u(I,\theta)$ are bounded and upper semicontinuous in (I,θ). Players share a common prior over states $F \in \Delta([0,1])$, where F has a compact support denoted by $\text{supp}(F) \equiv \Theta$.[13]

I impose three main assumptions on the base game. First, $u(I,\theta)$ is non-decreasing in I. Hence, an agent's incentive to invest (over not) is non-decreasing in the aggregate investment. Second, $v(I,\theta)$ is non-decreasing in I and $v(0,\theta) = 0$. Hence, the designer prefers more over less investment. Third, there exists sets $\overline{\Theta}, \underline{\Theta} \subseteq \Theta$ of strictly positive measure under F where for all $\theta \in \overline{\Theta}$, $u(0,\theta) > 0$ so investing is a strictly dominant action, and for all $\theta \in \underline{\Theta}$, $u(1,\theta) < 0$ so not investing is a strictly dominant action.

An *information structure* is a pair $\mathcal{S} \equiv (S, (\pi(\cdot|\theta))_{\theta\in\Theta})$.[14] $S \in \mathbb{B}(\mathbb{R})$ is a non-empty set of signals. Meanwhile, for each $\theta \in \Theta$, $\pi(\cdot|\theta) \in \Delta(\Delta(S))$ is a distribution over signal distributions $\mu \in \Delta(S)$. Denote the induced joint distribution over $S \times \Delta(S) \times \Theta$ by π, so $\pi(X \times Y \times Z) \equiv \int_Z \int_Y \int_X d\mu(s) d\pi(\mu|\theta) dF(\theta)$ for all $X \in \mathbb{B}(S)$, $Y \in \mathbb{B}(\Delta(S))$ and $Z \in \mathbb{B}(\Theta)$. Further let $\pi(\cdot|s)$ denote any version of the regular conditional probability on $\Delta(S) \times \Theta$ given signal $s \in S$ under π. I focus on $\mathcal{S}$ which induces a marginal distribution over signals with full support.

The designer first chooses an information structure $\mathcal{S}$. Then, the state $\theta \in \Theta$ is drawn according to F, and a signal distribution $\mu \in \Delta(S)$ is drawn according to $\pi(\cdot|\theta)$. Signals are then allocated *anonymously* across agents: each agent privately observes an independent draw of s from μ. By an appropriate "large of large numbers" (Sun, 2006), μ is the empirical distribution over other agents signals. Each agent then simultaneously chooses to invest or not, and payoffs are subsequently realized.

[13] Given a topological space X, $\mathbb{B}(X)$ denotes the Borel sigma-algebra of X. $\Delta(X)$ denotes the set of Borel probability measures over X. The support of a measure $\mu \in \Delta(X)$, denoted by $\text{supp}(\mu)$, is defined as the smallest closed set with measure one.

[14] This modelling approach is also used in Morris et al. (2022a) and Li et al. (2023).

3.2 Monotone Introspective Equilibrium

An information structure $\mathcal{S}$ induces a Bayesian game between agents. A (symmetric, pure) strategy for agents in this game is a map $\alpha : S \to \{0,1\}$. The aggregate investment under distribution μ given strategy α is $I(\mu|\alpha) \equiv \int_S \alpha(s)d\mu(s)$. A (symmetric) Bayes-Nash Equilibrium (BNE) of $\mathcal{S}$ is a strategy profile α^* under which no agent has a unilateral incentive to deviate. That is,

$$\alpha^*(s) \in \arg\max_{a\in\{0,1\}} \left\{ a \times \int_{\Delta(S)\times\Theta} u(I(\mu|\alpha^*),\theta)d\pi(\mu,\theta|s) \right\}, \quad \forall s \in S$$

The designer anticipates the BNE selected under an information structure to be its *monotone introspective equilibrium.* As in Section 2, this is the limit of an introspective best-response process, anchored on a *benchmark* of other agents' behaviours. Unlike Section 2, a benchmark is now an upper semicontinuous function $b : \Theta \to [0,1]$, where $b(\theta)$ is agents' benchmark about aggregate investment raised on state θ. For example, this captures the possibility that prior entrepreneurs had different success with raising investment for different quality projects, leading to different benchmarks.

Denote the set of benchmarks by $\mathcal{B}$. Definition 1 below formally defines the monotone introspective equilibrium of an information structure for a given benchmark.

Definition 1. *Given a benchmark $b \in \mathcal{B}$, let $(\alpha^{\mathcal{S},b,k})_{k\geq 1}$ denote the sequence of pure strategies defined as follows:*

1. *If $k = 1$, then*

$$\alpha^{\mathcal{S},b,1}(s) = \begin{cases} 1, & \int_{\Delta(S)\times\Theta} u(b(\theta),\theta)d\pi(\mu,\theta|s) \geq 0 \\ 0, & \int_{\Delta(S)\times\Theta} u(b(\theta),\theta)d\pi(\mu,\theta|s) < 0 \end{cases}, \quad \forall s \in S \tag{1}$$

2. *If $k > 1$, then*

$$\alpha^{\mathcal{S},b,k}(s) = \begin{cases} 1, & \int_{\Delta(S)\times\Theta} u(I(\mu|\alpha^{\mathcal{S},b,k-1}),\theta)d\pi(\mu,\theta|s) \geq 0 \\ 0, & \int_{\Delta(S)\times\Theta} u(I(\mu|\alpha^{\mathcal{S},b,k-1}),\theta)d\pi(\mu,\theta|s) < 0 \end{cases}, \quad \forall s \in S \tag{2}$$

If $(\alpha^{\mathcal{S},b,k})_{k\geq 1}$ is pointwise non-decreasing in k, then call the limit $\alpha^{\mathcal{S},b} \equiv \lim_{k\to\infty} \alpha^{\mathcal{S},b,k}$ the b-Monotone Introspective Equilibrium (b-MIE) of $\mathcal{S}$.

The next result says that b-MIE is indeed a BNE selection criterion. It also says that higher benchmarks[15] select larger BNE, which benefits the designer.

Lemma 1. *Take an information structure $\mathcal{S}$ and any two benchmarks $b, \tilde{b} \in \mathcal{B}$*

1. *If $\alpha^{\mathcal{S},b}$ exists, then it is a BNE of $\mathcal{S}$*
2. *If $b \geq \tilde{b}$, and both $\alpha^{\mathcal{S},b}$ and $\alpha^{\mathcal{S},\tilde{b}}$ exist, then $\alpha^{\mathcal{S},b}(s) \geq \alpha^{\mathcal{S},\tilde{b}}(s)$ for all $s \in S$.*

Part 1 holds as an agent's MIE strategy must be a best-response to other agents' strategies at high enough rounds of introspection and so in the limit. Part 2 holds because of the supermodularity of agents' payoffs. Higher benchmarks lead to more investment at the first round of introspection, which, by raising agents' payoffs from investing, leads to more investment at the second round, and so on.

3.3 Designer's Problem

Given benchmark b, let $\Pi(b)$ be the set of information structures with a b-MIE. $\Pi(b)$ is non-empty as it includes no disclosure.[16] Then, the designer's problem is to optimize over information structures in $\Pi(b)$, anticipating b-MIE selection:

$$V^*(b) \equiv \sup_{\mathcal{S}\in\Pi(b)} \int_{\Theta} \int_{\Delta(S)} v(I(\mu|\alpha^{\mathcal{S},b}), \theta) d\pi(\mu|\theta) dF(\theta) \tag{3}$$

The goal is to understand how changes in the benchmark affect optimal information provision, i.e., solutions to (3). In light of this, it is a priori unclear *how* to interpret the results obtained. This is because raising the benchmark affects not only the equilibrium selected $\alpha^{\mathcal{S},b}$, but also the feasible set of information structures $\Pi(b)$.

To resolve this, it is worth noting that the designer's payoff $V^*(b)$ is non-decreasing in the benchmark (see Section 5.5). Hence, if the designer (strictly) prefers to switch to choosing a new information structure after raising the benchmark, then it is because the BNE selected under the new information structure strictly outperforms that under the old one, not simply because the old information structure no longer has a MIE under the higher benchmark. Combining this observation with Lemma 1 yields the following interpretation for comparative statics in the benchmark.

[15] I say that b is higher than $\tilde{b}$, written as $b \geq \tilde{b}$ if $b(\theta) \geq \tilde{b}(\theta)$ for all $\theta \in \Theta$. A function $g : \mathcal{B} \to \mathbb{R}$ is non-decreasing in b if for all $b, \tilde{b} \in \mathcal{B}$, $b \geq \tilde{b}$ implies $g(b) \geq g(\tilde{b})$.

[16] I define no disclosure as any information structure with $|S| = 1$. It has a unique b-MIE in which all agents invest if $\int_\Theta u(b(\theta), \theta) dF(\theta) \geq 0$, and do not invest if $\int_\Theta u(b(\theta), \theta) dF(\theta) < 0$.

Remark 1. *A higher benchmark affects optimal information disclosure because it leads to less adversarial equilibrium selection.*

4 Introspective Implementation

This section builds the tools used to study the designer's problem under MIE. I first introduce *introspective outcomes.* These provide a possible description of agents' MIE behaviours under an information structure. I then fully characterize the set introspective outcomes implemented by some information structure, and provide a canonical class of information structures which implement them. Using these results, I then reformulate the designer's problem as one involving optimizing over introspective outcomes. Finally, I discuss the connection between the designer's problem studied here, to that under other selection criterion used in the literature.

4.1 Introspective Outcomes

Fix an information structure $\mathcal{S}$ with a b-MIE. An agent who observes signal $s \in S$ and invests in the b-MIE, so $\alpha^{\mathcal{S},b}(s) = 1$, can be divided into one of two roles. First, an agent can be an *anchor.* These are agents who prefer to invest against the benchmark and so switch at the first round, so $\alpha^{\mathcal{S},b,k}(s) = 1$ for all $k \geq 1$. Second, the agent can be a *non-anchor.* These are agents who do not invest against the benchmark, but eventually switches to invest, so $\alpha^{\mathcal{S},b,1}(s) = 0$ but $\alpha^{\mathcal{S},b}(s) = 1$. Observe then each signal distribution $\mu \in \Delta(S)$ can be associated to a pair $I^{\mathcal{S},b}(\mu) \equiv (I_A^{\mathcal{S},b}(\mu), I_N^{\mathcal{S},b}(\mu))$, where $I_A^{\mathcal{S},b}(\mu) \equiv \int_S \alpha^{\mathcal{S},b,1}(s) d\mu(s)$ is the mass of anchors induced, and $I_N^{\mathcal{S},b}(\mu) \equiv \int_S \alpha^{\mathcal{S},b}(s) d\mu(s) - I_A^{\mathcal{S},b}(\mu)$ is the mass of non-anchors induced.

Let $\mathcal{I} \equiv \{(I_A, I_N) \in [0,1]^2 : I_A + I_N \leq 1\}$ be the possible combinations of (masses of) anchors I_A and non-anchors I_N drawn. Call a map $\sigma : \Theta \to \Delta(\mathcal{I})$ an *introspective outcome*, where $\sigma(\cdot|\theta) \in \Delta(\mathcal{I})$ is the conditional distribution over masses of anchors and non-anchors on state θ. Because every information structure $\mathcal{S}$ (with a b-MIE) induces a distribution over signal distributions across states, it implements an introspective outcome σ, defined by

$$\sigma(W|\theta) \equiv \pi((I^{\mathcal{S},b})^{-1}(W)|\theta), \quad \forall W \in \mathbb{B}(\mathcal{I}),\ \forall \theta \in \Theta \tag{4}$$

Call an introspective outcome σ *b-implementable* if there exists an information structure $\mathcal{S}$ with a b-MIE such that (4) holds under σ. Note that the designer's payoff under any information structure which implements σ is the same, and is given by

$$V(\sigma) \equiv \int_{\Theta}\int_{\mathcal{I}} v(I_A + I_N, \theta) d\sigma(I_A, I_N|\theta) dF(\theta)$$

Finally, denote the total variation of (joint distribution induced by) an introspective outcome σ by[17]

$$\|\sigma\| \equiv \sup_{W \in \mathbb{B}(\mathcal{I}), \tilde{\Theta} \in \mathbb{B}(\mathcal{I})} \left| \int_{\tilde{\Theta}} \int_{W} d\sigma(I_A, I_N|\theta) dF(\theta) \right| \tag{5}$$

Then call an introspective outcome σ *approximately b-implementable* if there exists a sequence of b-implementable introspective outcomes $(\sigma^n)_{n\geq 1}$ that converges to σ in total variation. That is, σ can be approximately implemented by some information structure with a b-MIE.

4.2 Characterization of Implementability

I now characterize b-implementability. To start, the next result states a necessary condition for an introspective outcome to be b-implementable.

Theorem 1. *If an introspective outcome σ is b-implementable, then it is b-obedient. That is, it satisfies the following three conditions:*

1. ***Anchor obedience**:*

$$\int_{\Theta}\int_{\mathcal{I}} I_A u(b(\theta), \theta) d\sigma(I_A, I_N|\theta) dF(\theta) \geq 0 \tag{6}$$
$$\int_{\Theta}\int_{\mathcal{I}} I_A u(I_A, \theta) d\sigma(I_A, I_N|\theta) dF(\theta) \geq 0 \tag{7}$$

2. ***Non-anchor obedience**:*

$$\int_{\Theta}\int_{\mathcal{I}} \left(\int_{I_A}^{I_A+I_N} u(i, \theta) di \right) d\sigma(I_A, I_N|\theta) dF(\theta) \geq 0 \tag{8}$$

[17] By identifying an introspective outcome σ with the equivalence class $\{\sigma' : \|\sigma - \sigma'\| = 0\}$, I topologize the space of introspective outcomes with $\|\cdot\|$ moving forward.

3. ***Downwards obedience:***

$$\int_\Theta \int_\mathcal{I} I_N d\sigma(I_A, I_N|\theta)dF(\theta) > 0 \Rightarrow \int_\Theta \int_\mathcal{I} I_N u(b(\theta),\theta)d\sigma(I_A, I_N|\theta)dF(\theta) < 0 \tag{9}$$

$$\int_\Theta \int_\mathcal{I}(1 - I_A - I_N)u(b(\theta),\theta)d\sigma(I_A, I_N|\theta)dF(\theta) \leq 0 \tag{10}$$

$$\int_\Theta \int_\mathcal{I}(1 - I_A - I_N)u(I_A + I_N,\theta)d\sigma(I_A, I_N|\theta)dF(\theta) \leq 0 \tag{11}$$

The intuition is as follows. Take an information structure $\mathcal{S}$ that b-implements the introspective outcome σ. Equations (6) and (7) are the aggregate payoffs of anchors from investing against, respectively, the benchmark and when other agents follow $\alpha^{\mathcal{S},b,1}(\cdot)$. Thus, both are positive. (8) is obtained from aggregating the payoffs of non-anchors from investing at the earliest round of introspection in which the agent switches to invest, $k = \min\{k' : \alpha^{\mathcal{S},b,k'}(s) = 1\}$. Thus, it is positive. (9) is the aggregate payoff of non-anchors from investing against the benchmark, which is negative. Finally (10) and (11) are the aggregate payoffs of agents who do not invest under the signal observed, from investing against, respectively, the benchmark and assuming only other anchors and non-anchors invest. Thus, both are negative.

The next result says that b-obedience is basically sufficient for b-implementability.

Theorem 2. *If an introspective outcome σ is b-obedient, then it is approximately b-implementable.*

The proof involves constructing, for any b-obedient introspective outcome σ, a sequence of information structures which implement introspective outcomes that approximate σ. These information structures share two key features. First, they supply (approximately) all anchors symmetric information. This is because the main requirement to be an anchor, to prefer to invest against the benchmark, is symmetric across agents. Second, different non-anchors are supplied different information. This is done to maximize the heterogeneity in the rounds non-anchors switch to invest during introspection. Because agents' payoffs are increasing in investment, this maximizes their aggregate incentives to switch to invest. Hence, the choice of an b-obedient introspective outcome can be understood as choosing the distribution over symmetric (anchors) and asymmetric information (non-anchors) supplied to agents to induce investment. I use this interpretation moving forward.

Figure 5 gives an example of a constructed information structure that implements an introspective outcome which approximates σ. For simplicity, I assume here that σ

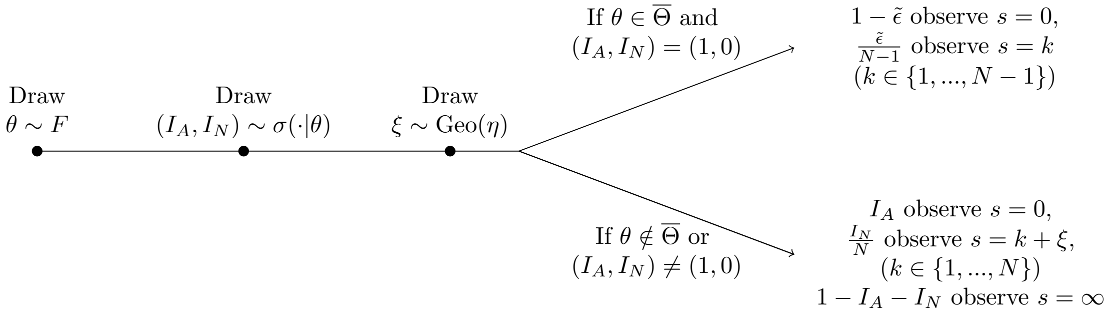

Draw
$\theta \sim F$
Draw
$(I_A, I_N) \sim \sigma(\cdot|\theta)$
Draw
$\xi \sim \text{Geo}(\eta)$
If $\theta \in \overline{\Theta}$ and
$(I_A, I_N) = (1, 0)$
$1 - \tilde{\epsilon}$ observe $s = 0$,
$\frac{\tilde{\epsilon}}{N-1}$ observe $s = k$
$(k \in \{1, ..., N - 1\})$
If $\theta \notin \overline{\Theta}$ or
$(I_A, I_N) \neq (1, 0)$
I_A observe $s = 0$,
$\frac{I_N}{N}$ observe $s = k + \xi$,
$(k \in \{1, ..., N\})$
$1 - I_A - I_N$ observe $s = \infty$

Figure 5: An information structure that approximately implements σ. Here, $0 < \tilde{\epsilon}$ is taken sufficiently small, $0 < \eta \ll \tilde{\epsilon}$, and $\text{Geo}(\eta)$ denotes the geometric distribution with parameter η

also satisfies b-obedience strictly, and draws $(I_A, I_N) = (1, 0)$ on all states $\theta \in \overline{\Theta}$ with probability at least $\epsilon > 0$, where investing is a strictly dominant action for agents. Also, $N > 1$ is large enough so that non-anchor obedience in (8) still holds strictly when replacing $\int_{I_A}^{I_A+I_N} u(i, \theta)di$ with $\sum_{i=1}^{N} u(I_A + I_N(i-1)/N, \theta)/N$.

Note that agents observing $s \leq N - 1$ are anchors, $N \leq s < \infty$ are non-anchors, and $s = \infty$ do not invest. To see this, first observe that as $\tilde{\epsilon}$ is small, I_A agents observe $s = 0$ with probability $\approx \sigma(I_A, I_N|\theta)$ on all states θ. Since σ satisfies anchor obedience strictly, these agents' expected payoffs from investing against the benchmark and when only other agents are positive. In turn, all such agents invest at the first two rounds of introspection which ensures they invest at all higher rounds as anchors. A similar argument implies all agents observing $s = \infty$ do not invest. Finally, as $\tilde{\epsilon} \gg \eta$, an agent observes $1 \leq s \leq N - 1$ with large probability on states in $\overline{\Theta}$. Hence, they find investing a strictly dominant action, and so invest at all rounds of introspection as an anchor. Meanwhile, any agent observes $s \geq N$ with frequencies close to $\int_{\mathcal{I}} I_N d\sigma(I_A, I_N|\theta)$ on state θ. Since σ satisfies downwards obedience in (9) strictly, all such agents do not invest at the first round of introspection. Meanwhile, the payoff from investing provided lower signal agents invest is approximately the expectation of $\sum_{i=1}^{N} u(I_A + I_N(i-1)/N, \theta)/N$ under σ. As $\int_{I_A}^{I_A+I_N} u(i, \theta)di \approx \sum_{i=1}^{N} u(I_A + I_N(i-1)/N, \theta)/N$ and σ satisfies non-anchor obedience strictly, this is positive. An induction argument on N then implies all such agents switch to investing at some round larger than s. Thus, these agents are non-anchors.

The above is important as for any pair (I_A, I_N) drawn, if $\xi \geq N$ is also drawn, then I_A agents observe $s \leq N - 1$, i.e., are anchors, and I_N agents observe $N \leq s < \infty$, i.e., are non-anchors. Since $\xi \geq N$ is drawn with probability ≈ 1 for $\tilde{\epsilon} \approx 0$, the introspective outcome implemented by Figure 5 approximates σ. Furthermore. the fraction of anchors observing $s = 0$ goes to one as $\tilde{\epsilon} \to 0$. Meanwhile, as $N \to \infty$, the mass of non-anchors observing a single $s \geq N$ goes to zero. In turn, the mass of agents observing symmetric signals is approximately the mass of anchors.

4.3 Information Design as Introspection Manipulation

I now use Theorems 1 and 2 to reformulate the designer's problem to one involving optimizing over introspective outcomes. First, combining the two results implies

$$\{\sigma : \sigma \text{ is } b\text{-implementable}\} \subseteq \{\sigma : \sigma \text{ is } b\text{-obedient}\} \subseteq \left\{\sigma : \begin{array}{c} \sigma \text{ is approximately} \\ b\text{-implementable} \end{array}\right\}$$

The right-most set is the closure of the left-most set. Furthermore, the designer's payoff as a function of introspective outcomes $V(\cdot)$ is continuous.[18] Therefore, the designer's payoff in the problem (3), $V^*(b)$, can be identified by optimizing over approximately b-implementable introspective outcomes. But this can also be achieved by optimizing over b-obedient introspective outcomes. Hence,

$$V^*(b) = \sup_{\sigma \text{ is } b\text{-obedient}} V(\sigma) \tag{12}$$

Next, I will simplify (12) further. Say that an introspective outcome σ *investment dominates* (resp. *anchor dominates*) another one σ' if the distribution over aggregate investment (resp. anchors) under σ first-order stochastically dominates the one induced by σ'.[19] As the designer's payoff is monotone in investment, if σ investment dominates σ', then the designer weakly prefers σ over σ'. Meanwhile, by the discussion in Section 4.2, if σ anchor dominates σ', then the designer supplies more symmetric information to induce investment under σ over σ'.

Say an introspective outcome is *upper b-obedient* if it satisfies anchor obedience (6) and (7), and non-anchor obedience (8), but not necessarily downwards obedience. Thus, these are characterized by fewer obedience constraints than b-obedient intro-

[18] By definition, $V(\sigma)$ is linear in σ. Furthermore, for all σ, $|V(\sigma)| \leq \max_{\theta \in \Theta} v(1, \theta) \, \|\sigma\|$. Hence, $V(\cdot)$ is a bounded and so continuous linear functional.

[19] That is, σ investment dominates (resp. anchor dominates) σ' if for all non-decreasing functions $f : [0, 1] \to \mathbb{R}$ and all states $\theta \in \Theta$, $\int_{\mathcal{I}} f(I_A + I_N) d\sigma(I_A, I_N|\theta) \geq \int_{\mathcal{I}} f(I_A + I_N) d\sigma'(I_A, I_N|\theta)$ (resp. $\int_{\mathcal{I}} f(I_A) d\sigma(I_A, I_N|\theta) \geq \int_{\mathcal{I}} f(I_A) d\sigma'(I_A, I_N|\theta)$).

spective outcomes. The next result simplifies the designer's to optimizing over upper b-obedient introspective outcomes.

Proposition 1. *Take an upper b-obedient introspective outcome σ. Then, there exists a b-obedient introspective outcome that investment and anchor dominates σ. Hence,*

$$V^*(b) = \sup_{\sigma \text{ is upper } b\text{-obedient}} V(\sigma) \tag{P}$$

The proof shows that a two step modification can be applied to any upper b-obedient introspective outcome σ to make it b-obedient. Step 1 involves, for each pair (I_A, I_N) drawn, raising the total mass of anchors and non-anchors to one, unless both $u(b(\theta), \theta) < 0$ and $u(I_A, I_N, \theta) < 0$ hold. This means that any time there are agents who do not invest, their expected payoff from investing against the benchmark and when all other anchors and non-anchors invest are both negative. Hence, the result satisfies conditions (10) and (11) for downwards obedience. Step 2 involves pooling the entire mass of non-anchors to anchors if (9) for downwards obedience is violated. Hence, the result satisfies b-obedience. Further notice every step weakly raises the aggregate investment and anchors induced, so the result investment and anchor dominates σ.

Moving forward, I refer to (P) as "the" designer's problem, and call any solution to (P) an *optimal* introspective outcome. The next result shows that a solution exists.

Lemma 2. *There exists an optimal introspective outcome σ.*

As the designer's payoff is monotone in investment, an optimal introspective outcome is generally b-obedient. If so, then it describes the distribution over roles that the designer prefers to implement and, by the discussion in Section 4.2, the type of information the designer can supply to implement it. Even if it is not b-obedient, it still identifies a lower bound on the amount of investment and anchors induced under a b-obedient introspective outcome that the designer equally prefers to it, which can be constructed via the procedure described following Proposition 1. Thus, it is without loss of generality to study optimal information disclosure through problem (P).

4.4 Connection to Other Selection Criteria

I conclude by discussing the connection between the designer's problem (P), to the designer's problem under other equilibrium selection criteria. All proofs and additional details in this section are given in the supplementary appendix.

Extreme selection criterion. I first discuss the connection to the designer's problem under the extreme designer-preferred and *adversarial* equilibrium selection. Both are used extensively in the information design literature.

Call a map $q : \Theta \to \Delta([0,1])$ an *outcome*, where $q(\cdot|\theta) \in \Delta([0,1])$ represents the distribution over the aggregate investment of agents under q on state θ. Let $\|q\|$ denote the total variation of (the joint distribution induced by) the outcome q, extended appropriately from (5).[20]

Let $\overline{\alpha}^{\mathcal{S}}$ and $\underline{\alpha}^{\mathcal{S}}$ denote, respectively, the largest and smallest BNE of an information structure $\mathcal{S}$. These are, respectively, a designer-preferred and adversarial BNE of $\mathcal{S}$.[21] Say that an outcome q is designer-preferred *implementable* if it is the designer-preferred equilibrium distribution over investment of some information structure. That is, if there exists an information structure $\mathcal{S}$ such that

$$q^n(W|\theta) = \pi(I^{-1}(W|\overline{\alpha}^{\mathcal{S}})|\theta) \quad \forall W \in \mathbb{B}([0,1]),\ \theta \in \Theta \tag{13}$$

Call an outcome q *approximately designer-preferred implementable* if there exists a sequence of designer-preferred implementable outcomes $(q^n)_{n\geq 1}$ that converges to q in total variation. Approximate implementability for adversarial selection and b-MIE are defined verbatim, replacing $\overline{\alpha}^{\mathcal{S}}$ in (13) with $\underline{\alpha}^{\mathcal{S}}$ and $\alpha^{\mathcal{S},b}$ respectively.

In the standard approach to optimal information design, one optimizes over (approximately) implementable outcomes under the selection criterion, and then backs out an information structure which (approximately) implements the solution through an appropriate "revelation principle" (Bergemann and Morris, 2019). In the case of designer-preferred and adversarial selection, these problems are given by, respectively,

$$\overline{V} \equiv \sup_{\substack{q \text{ is approximately} \\ \text{designer-preferred implementable}}} \int_{\Theta} \int_0^1 v(I,\theta) dq(I|\theta) dF(\theta)$$

$$\underline{V} \equiv \sup_{\substack{q \text{ is approximately} \\ \text{adversarially implementable}}} \int_{\Theta} \int_0^1 v(I,\theta) dq(I|\theta) dF(\theta)$$

Similarly, an alternative, albeit less useful, formulation of the designer's problem (P)

[20] Formally, $\|q\| \equiv \sup_{W\in\mathbb{B}([0,1]),\tilde{\Theta}\in\mathbb{B}(\mathcal{I})} |\int_{\tilde{\Theta}} \int_W dq(I|\theta)dF(\theta)|$.

[21] As an agent's payoff u may not be lower semicontinuous, for a non-generic information structure $\mathcal{S}$, $\underline{\alpha}^{\mathcal{S}}$ may not exist. If so, then to avoid this minor technicality, I let $\underline{\alpha}^{\mathcal{S}}$ be the **0**-MIE of $\mathcal{S}$, which always exists. As I discuss later, the two coincide for generic $\mathcal{S}$, which justifies this approach.

is as optimizing over approximately b-MIE implementable outcomes. As such, one way of connecting the designer's problem under extreme selection criterion to b-MIE is to connect the outcomes approximately implementable under each. Here, it is useful to recall that higher benchmarks capture less adversarial equilibrium selection under MIE (Remark 1). Proposition 2 below shows that this means outcomes implementable under extreme selection criterion are connected to those implementable under the extreme benchmarks $\mathbf{0}$ and $\mathbf{1}$ taking on, respectively, a value of 0 and 1 on all states.

Proposition 2.

1. *An outcome is approximately adversarially implementable if and only if it is approximately* $\mathbf{0}$*-MIE implementable. Hence,* $V^*(\mathbf{0}) = \underline{V}$*.*

2. *If an outcome is approximately* $\mathbf{1}$*-MIE implementable, then it is approximately designer-preferred implementable. If an outcome is approximately designer-preferred implementable, then there exists an approximately* $\mathbf{1}$*-MIE implementable outcome that first-order stochastically dominates it.*[22] *Hence,* $V^*(\mathbf{1}) = \overline{V}$*.*

For Part 1, first note that the adversarial BNE of an information structure is the limit of the pure strategies sequence in Definition 1, but breaking ties in favour of not investing (Milgrom and Roberts, 1990). For generic information structures, this *is* the $\mathbf{0}$-MIE. In fact, I find that any approximately implementable outcome under either $\mathbf{0}$-MIE or adversarial selection can be approximated by outcomes implementable by information structures where the $\mathbf{0}$-MIE is the adversarial equilibrium. This implies the designer's problem under $\mathbf{0}$-MIE and adversarial selection coincide.[23]

For Part 2, first note that under designer-preferred selection, all optimal outcomes for the designer are approximately implementable by direct information structures, where agents observe "recommendations" to invest or not invest. These are simply a special case of the information structures discussed in Section 4.2 which never draw signals $1 \leq s < \infty$, and where one interprets signal $s = 0$ and $s = \infty$ to invest and not to invest respectively. I also find that these "anchor-only" information structures support all outcomes implementable under $\mathbf{1}$-MIE. This is because under

[22] Given outcomes q, q', q *first-order stochastically dominates* q' if for all $\theta \in \Theta$ and all non-decreasing functions $f : [0,1] \to \mathbb{R}$, $\int_0^1 f(I)dq(I|\theta) \geq \int_0^1 f(I)dq'(I|\theta)$.

[23] Another way to see the connection is as follows. Consider the subset of information structures described in Section 4.2 which never draw signal $s = 0$. Morris et al. (2024) shows that (a finite agent analogue of) such information structures support all outcomes implementable under adversarial selection. I show that the same is true for outcomes approximately implementable under $\mathbf{0}$-MIE.

the highest benchmark $\mathbf{1}$, anchors prefer to (switch to) invest against the benchmark so long as investing is not strictly dominated, so inducing investment via anchors is always "less difficult" than non-anchors. This yields Part 2. In fact, when one uses designer-preferred selection, it is standard to *only* focus on optimizing over outcomes implementable by direct information structures (Bergemann and Morris, 2019). Thus, the designer's problem under $\mathbf{1}$-MIE and designer-preferred selection coincide.

Other selection criterion. The next result provides conditions for when the payoff for the designer under *every* selection criterion is attained under b-MIE for some benchmark b. This speaks to the richness of MIE as a selection criterion.

Proposition 3. *Suppose u and v are continuous. Then for any payoff for the designer attained under a selection criterion $V \in [\underline{V}, \overline{V}]$, there exists a benchmark $b \in \mathcal{B}$ such that $V = V^*(b)$.*

Proposition 3 holds because $V^*(b)$ is continuous in the class of constant benchmarks $\{\lambda \mathbf{0} + (1-\lambda)\mathbf{1}\}_{\lambda \in [0,1]}$ under the stated assumptions. To establish this, I show in the supplementary appendix that the set of joint distributions over $\mathcal{I} \times \Theta$ induced by upper b-obedient introspective outcomes is weak*-compact[24] and continuous in constant benchmarks. Likewise, the designer's objective $V(\cdot)$ (adapted to joint distributions) is weak* continuous. Applying Berge's Maximum Theorem then yields the claim. One example for when this holds is the setting of Section 2.

5 Optimal Information Design

In this section, I study the designer's problem for the class of *threshold games.*

Assumption 1 (Threshold Games). *For all $\theta \in \Theta$, and all $I \in \{I' \in [0,1] : u(I', \theta) < 0\}$, $u(I, \theta) = u(0, \theta)$.*

In a threshold game, agents payoffs are constant from a marginal increase in aggregate investment unless a threshold is met, i.e., $I \geq \underline{I}(\theta) \equiv \sup\{I' \in [0,1] : u(I', \theta) < 0\}$. These games model many relevant economic settings. Two are discussed next.

[24] This is why the joint distribution approach is needed. By Riesz's Lemma, the set of upper b-introspective outcomes is not $\|\cdot\|$-compact, which precludes the use of Berge's Theorem.

Example 1. *Suppose there exists a constant $c > 0$ and upper semicontinuous functions $\tilde{I} : \Theta \to \mathbb{R}$ and $W : \Theta \to (c, \infty)$ such that $u(I, \theta) = W(\theta) - c > 0$ if $I \geq \tilde{I}(\theta)$, and $u(I, \theta) - c$ if $I < \tilde{I}(\theta)$. Then, the base game is a game of regime change: the regime is maintained if and only if there is sufficient investment ($I \geq \tilde{I}(\theta)$), and agents benefit only when the regime is maintained. This models many settings, including team project coordination in organizations and bank-runs (see Morris and Shin, 2003). Recent papers studying information design in this setting include Goldstein and Huang (2016), Li et al. (2023) and Inostroza and Pavan (2023).*

Example 2. *Suppose that on all states $\theta \in \Theta$, either investing or not investing is a dominant action, so $u(0, \theta) \geq 0$ or $u(1, \theta) \leq 0$ holds. Furthermore, if not investing is dominant, so $u(1, \theta) \leq 0$, then agents do not benefit from higher aggregate investment, so $u(0, \theta) = u(1, \theta)$. This game generalizes the setting in Section 2. It is also a "one-sided" counterpart to the investment game studied in Bergemann and Morris (2016, 2019) and Mathevet et al. (2020), the main difference being that there are only positive complementarities on states in which investing is a dominant action.*

I also impose the following two minor assumptions throughout.

Assumption 2. *F is continuous with support $\Theta = [0, 1]$.*

Assumption 3. *There exists a $\theta_0 \in [0, 1]$ such that if $v(0, \theta) = v(1, \theta)$ and $\int_0^1 u(i, \theta)di < 0$, then $\theta \leq \theta_0$.*

Assumption 2 will imply the existence of a *deterministic* optimal introspective outcome, that draws a unique mass of anchors and non-anchors on each state. Assumption 3 labels states so that the smallest states are those in which the designer and non-anchors both prefer not to have any investment. Both help to simplify the statement of the results, but are otherwise not necessary for the main insights to hold.

This section proceeds as follows. I begin by solving a relaxed version of the designer's problem. I then use the solution to the relaxed problem to construct a selection of optimal introspective outcomes, one for each benchmark. Among these, I characterize how changes in the benchmark affect three key quantities: the amount of aggregate investment and mass of anchors induced across states, the designer's incentive to perfectly coordinate investments, and players' payoffs. Finally, I discuss how several of these results can be extended to non-threshold games.

5.1 Relaxed Problem

Relaxed problem. To motivate the relaxed problem, I first discuss the notion of *perceived payoffs* from investment. This is the payoff an agent believes he obtains from investing, at the round of introspection in which the agent switches to investing. For an anchor, this is $u(b(\theta), \theta)$. For a non-anchor, this is $\int_{I_A}^{I_A+I_N} u(i, \theta)di/I_N$. The total perceived payoffs of I_A anchors and I_N non-anchors from investing on state θ is

$$U(I_A, I_N|\theta) \equiv I_A u(b(\theta), \theta) + \int_{I_A}^{I_A+I_N} u(i, \theta)di$$

A necessary condition for an introspective outcome to be upper b-obedient is for the expectation of the total perceived payoffs of agents from investing to be non-negative. That is, for the sum of the first anchor obedience constraint (6), which captures anchors' expected perceived payoffs from investment, and non-anchor obedience (8), which captures non-anchors' expected perceived payoffs from investment, to be non-negative. This leads to following natural relaxation of the designer's problem:

$$\max_{\sigma} \int_{\Theta} \int_{\mathcal{I}} v(I_A, I_N, \theta) d\sigma(I_A, I_N|\theta) dF(\theta) \tag{R}$$

$$\text{s.t.} \quad \int_{\Theta} \int_{\mathcal{I}} U(I_A, I_N|b, \theta) d\sigma(I_A, I_N|\theta) dF(\theta) \geq 0 \tag{14}$$

I call any introspective outcome which satisfies (14) *feasible.*

Notice that the objective function and constraints are linear in the introspective outcome. Thus, the relaxed problem (R) is a linear programming problem. This motivates the use of a duality approach to solving (R), which I introduce next.

Dual problem. The *dual problem* to (R) involves choosing a multiplier $\lambda \geq 0$ and a measurable function $\phi : \Theta \to \mathbb{R}$ to solve

$$\min_{\lambda,\phi} \int_{\Theta} \phi(\theta) dF(\theta)$$

$$\text{s.t.} \quad \phi(\theta) \geq \mathcal{L}(I_A, I_N|\lambda, b, \theta), \quad \forall((I_A, I_N), \theta) \in \mathcal{I} \times \Theta \tag{15}$$

where $\mathcal{L}(\cdot|\lambda, b, \theta)$, defined below, is the state-wise *Lagrangian*

$$\mathcal{L}(I_A, I_N|\lambda, b, \theta) \equiv v(I_A + I_N, \theta) + \lambda U(I_A, I_N|b, \theta), \quad \forall((I_A, I_N), \theta) \in \mathcal{I} \times \Theta$$

By inspecting the constraint (15), it holds that if (λ, ϕ) solves the dual problem, then so does $(\lambda, \phi^b(\cdot|\lambda))$, where

$$\phi^b(\theta|\lambda) \equiv \max_{(I_A, I_N)\in\mathcal{I}} \mathcal{L}(I_A, I_N|\lambda, b, \theta), \quad \forall\theta \in \Theta$$

Hence, the dual problem can be restated as solving

$$V^D(b) \equiv \min_{\lambda\geq 0} \int_\Theta \max_{(I_A,I_N)\in\mathcal{I}} \mathcal{L}(I_A, I_N|\lambda, b, \theta) dF(\theta) \tag{D}$$

Solving the relaxed problem. Let λ^b denote the smallest solution to (D). I will use λ^b to solve the relaxed problem (R). To do so, I make use of two technical results. The first, *complementary slackness*, provides a necessary and sufficient condition for a feasible introspective outcome to solve the relaxed problem.

Lemma 3. *A feasible introspective outcome σ solves (R) if and only if*

$$\mathrm{supp}(\sigma(\cdot|\theta)) \subseteq \max_{(I_A,I_N)\in\mathcal{I}} \mathcal{L}(I_A, I_N|\lambda^b, b, \theta) \text{ for almost all } \theta \in \Theta, \quad \text{and} \tag{C1}$$

$$\lambda^b \int_\Theta \int_\mathcal{I} U(I_A, I_N|b, \theta) d\sigma(I_A, I_N|\theta) dF(\theta) = 0 \tag{C2}$$

The second result identifies a sufficient condition on the support of an introspective outcome on each state for (C1) to hold. Notably, this condition varies depending on whether the state is an *agreement state*, i.e., whether $u(b(\theta), \theta) \geq 0$ holds. Denote the set of b-agreement states by $\overline{\Theta}^b$.

Lemma 4. *Take an introspective outcome σ. If*

1. *For all $\theta \in \overline{\Theta}^b$, $\mathrm{supp}(\sigma(\cdot|\theta)) = \{(b(\theta), 1 - b(\theta))\}$, and*
2. *For all $\theta \notin \overline{\Theta}^b$, $\mathrm{supp}(\sigma(\cdot|\theta)) \subseteq \{(0, I) : I \in \max_{I'\in[0,1]} \mathcal{L}(0, I'|\lambda, b, \theta)\}$,*

then σ satisfies (C1).

I now use Lemma 3 and 4 to construct a solution to the relaxed problem (R). Denote[25]

$$I^b_-(\theta) \equiv \min \arg\max_{I\in[0,1]} \mathcal{L}(0, I|\lambda^b, b, \theta), \qquad I^b_+(\theta) \equiv \max \arg\max_{I\in[0,1]} \mathcal{L}(0, I|\lambda^b, b, \theta)$$

[25] These exist as $v(I, \theta)$ is upper semicontinuous in I, so $\arg\max_{I\in[0,1]} \mathcal{L}(0, I|\lambda^b, \theta)$ is compact.

Consider the class of introspective outcomes $\{\underline{\sigma}^b_{\underline{\theta}}\}_{\underline{\theta}\in[0,1]}$ defined as follows:

$$\underline{\sigma}^b_{\underline{\theta}}(\cdot|\theta) \equiv \begin{cases} \delta_{(b(\theta),1-b(\theta))}, & \theta \in \overline{\Theta}^b \\ \delta_{(0,I^b_+(\theta))}, & \theta \geq \underline{\theta} \text{ and } \theta \notin \overline{\Theta}^b \\ \delta_{(0,I^b_-(\theta))}, & \theta < \underline{\theta} \text{ and } \theta \notin \overline{\Theta}^b \end{cases}$$

Every $\underline{\sigma}^b_{\underline{\theta}}$ draws $b(\theta)$ anchors and $1-b(\theta)$ non-anchors on all b-agreement states, and induces 0 anchors and smallest or largest number of non-anchors that maximizes $\mathcal{L}(0,I|\lambda,b,\theta)$ on all non b-agreement states. By Lemma 4, any $\underline{\sigma}^b_{\underline{\theta}}$ satisfies (C1). Therefore, if $\underline{\sigma}^b_{\underline{\theta}}$ is also feasible and satisfies condition (C2), then Lemma 3 says it solves the relaxed problem. The proof of Proposition 4 below shows that both of these are satisfied under $\underline{\sigma}^b \equiv \underline{\sigma}^b_{\underline{\theta}^b}$ where

$$\underline{\theta}^b \equiv \sup\left\{\underline{\theta}\in[0,1] : \int_\Theta\int_{\mathcal{I}} U(I_A,I_N|b,\theta)\underline{\sigma}^b_{\underline{\theta}}(I_A,I_N|\theta)dF(\theta) \geq 0\right\} \tag{16}$$

This yields the following.

Proposition 4. *The introspective outcome $\underline{\sigma}^b$ solves the relaxed problem (R).*

5.2 Optimal Introspective Outcome

I will now use $\underline{\sigma}^b$ to construct a solution to the designer's problem. Denote the aggregate investment induced on state θ under $\underline{\sigma}^b$ by $I^b(\theta) \equiv \int_{\mathcal{I}}(I_A+I_N)d\underline{\sigma}^b(I_A,I_N|\theta)$. Consider the class of introspective outcomes $\{\sigma^b_{\overline{\theta}}\}_{\overline{\theta}\in[0,1]}$ defined by

$$\sigma^b_{\overline{\theta}}(\cdot|\theta) \equiv \begin{cases} \delta_{(b(\theta),1-b(\theta))}, & \theta\in\overline{\Theta}^b \\ \delta_{(\min\{\underline{I}(\theta),I^b(\theta)\},\max\{0,I^b(\theta)-\underline{I}(\theta)\})}, & \theta \geq \overline{\theta} \text{ and } \theta\notin\overline{\Theta}^b \\ \delta_{(0,I^b(\theta))}, & \theta < \overline{\theta} \text{ and } \theta \notin \overline{\Theta}^b \end{cases} \tag{17}$$

Each $\sigma^b_{\overline{\theta}}$ induces the same aggregate investment across states as $\underline{\sigma}^b$, so the designer is indifferent between $\sigma^b_{\overline{\theta}}$ and $\underline{\sigma}^b$. However, each $\sigma^b_{\overline{\theta}}$ raises the mass of anchors and lowers the mass of non-anchors on non b-agreement states $[\overline{\theta},1]\backslash\overline{\Theta}^b$. This lowers anchors' perceived payoffs from investment, and raises non-anchors' perceived payoffs from investment by, respectively,

$$\int_{[\overline{\theta},1]\setminus\overline{\Theta}^b} \min\{\underline{I}(\theta), I^b(\theta)\}u(b(\theta),\theta)dF(\theta) = \int_{[\overline{\theta},1]\setminus\overline{\Theta}^b} \min\{\underline{I}(\theta), I^b(\theta)\}u(0,\theta)dF(\theta) \leq 0 \tag{18}$$

$$\int_{[\overline{\theta},1]\setminus\overline{\Theta}^b} \int_0^{\min\{\underline{I}(\theta), I^b(\theta)\}} u(i,\theta)dF(\theta) = -\int_{[\overline{\theta},1]\setminus\overline{\Theta}^b} \min\{\underline{I}(\theta), I^b(\theta)\}u(0,\theta)dF(\theta) \geq 0 \tag{19}$$

Importantly, the right-hand sides of (18) and (19) sum to zero.[26] This means that lowering the threshold state $\overline{\theta}$ transfers the perceived "cost" of investment from non-anchors to anchors. It also means that the total perceived payoffs of anchors and non-anchors from investing under $\sigma^b_{\overline{\theta}}$ is the same as that under the relaxed problem solution. $\underline{\sigma}^b$, which is positive. Therefore, lowering $\overline{\theta}$ as much as possible while ensuring anchors' perceived payoffs from investing are non-negative, i.e., until

$$\overline{\theta}^b \equiv \min\left\{\overline{\theta} \in [0,1] : \int_{\overline{\Theta}^b} b(\theta)u(b(\theta),\theta)dF(\theta) + \int_{[\overline{\theta},1]\setminus\overline{\Theta}^b} \min\{\underline{I}(\theta), I^b(\theta)\}u(0,\theta)dF(\theta) \geq 0\right\}$$

will also ensure non-anchors' perceived payoffs from investing are non-negative. I show in the Appendix that resulting introspective outcome $\sigma^b \equiv \sigma^b_{\overline{\theta}^b}$ will also be upper b-obedient. Thus, it solves the designer's problem.

Proposition 5. *σ^b is an optimal introspective outcome.*

The main insight is that for any threshold game, an optimal introspective outcome can be constructed via a simple two-step approach. First, solve the relaxed problem to pin down agents' aggregate investment on all states, and the mass of anchors induced on agreement states. Second, transfer slack from anchors to non-anchors by raising the mass of anchors on non agreement states. The next example shows how to use this procedure to solve the designer's problem in Section 2.

Example 3. *First, solving the relaxed problem yields the introspective outcome $\underline{\sigma}^b$ with $\underline{\sigma}^b((b,1-b)|H) = 1$, $\underline{\sigma}^b((0,1)|L) = (3+b^2)/4$ and $\underline{\sigma}^b((0,0)|L) = (1-b^2)/4$. Under it, anchors' and non-anchors' perceived payoffs from investment are $b(1+b)/4 > 0$ and $-b(1+b)/4 < 0$ respectively. To fix this, replace the pair $(0,1)$ drawn on state L with $(2b(1+b)/(3+b^2), (3-b^2-2b)/(3+b^2))$, i.e., induce less non-anchors and more anchors. This yields the optimal introspective outcome described by Figure 3.*

[26] This holds as in a threshold game, $u(b(\theta),\theta) = u(0,\theta)$ on all non b-agreement states (so the equality in (18) holds), while $u(I,\theta) = u(0,\theta)$ for all $I < \underline{I}(\theta)$ (so the equality in (19) holds).

Moving forward, I refer to σ^b as "the" optimal introspective outcome. While there can be others, Lemma 3 implies that under any optimal introspective outcome, on almost all b-agreement states, any pair (I_A, I_N) drawn satisfies $v(I_A + I_N, \theta) = v(1, \theta)$ and $u(I_A, \theta) = u(b(\theta), \theta)$, while on almost all non b-agreement states, the aggregate investment induced satisfies $I \in [I^b_-(\theta), I^b_+(\theta)]$. As the next example shows, this means that σ^b often captures agents' behaviours under all optimal introspective outcomes.

Example 4. *Consider again the example of Section 2, where I label the introspective outcome in Figure 3 by σ^b. Since $v(I, \theta)$ and $u(I, \theta)$ are strictly increasing in I on the agreement state $\theta = H$, like σ^b, all optimal introspective outcomes must drawn $b(\theta)$ anchors and $1 - b(\theta)$ non-anchors on H with probability one. Meanwhile, the mass of anchors and non-anchors induced on L must ensure anchor and non-anchor obedience bind. This implies all optimal introspective outcomes induce the same expected number of anchors and non-anchors (and so aggregate investment) on state L as σ^b.*

5.3 Leveraging the Benchmark

I now discuss how the designer leverages an increase in the benchmark, i.e., by changing the optimal introspective outcome σ^b, to raise her payoff. The first result connects the benchmark to the aggregate investment on each state under σ^b.

Proposition 6. *Take any two benchmarks $b, \tilde{b}$. If $b \geq \tilde{b}$, then for all $\theta \in \Theta$, $I^b(\theta) \geq I^{\tilde{b}}(\theta)$. Hence, raising the benchmark raises aggregate investment on all states.*

The intuition is as follows. Consider the benchmark $\tilde{b}$. Under the optimal introspective outcome $\sigma^{\tilde{b}}$, agents' total perceived payoffs from investing are positive on all $\tilde{b}$-agreement states, and (often) negative on non $\tilde{b}$-agreement states. That is, the designer "spends" the incentives generated for agents to invest on $\tilde{b}$-agreement states, to induce investment on non $\tilde{b}$-agreement states. Raising the benchmark to $b \geq \tilde{b}$ not only expands the set of b-agreement states, on which all agents invest, but also raises agents' total perceived benefit from investing on all such states. This gives the designer more to "spend" on inducing investment on non b-agreement states. Because the designer's payoff and obedience constraints are additively separable across states, the designer then optimally raises investment on all non b-agreement states.

The next result characterizes how the mass of anchors on each state under the optimal introspective outcome σ^b, $I^b_A(\theta) \equiv \int_{\mathcal{I}} I_A d\sigma^b(I_A, I_N|\theta)$, varies in the benchmark.

Here, recall by the discussion in Section 4.2 that the mass of anchors also captures the amount of symmetric information provided to agents to induce investment.

Proposition 7. *Take any two benchmarks $b, \tilde{b}$. If $b \geq \tilde{b}$, then on all $\theta \in \overline{\Theta}^b$, $I_A^b(\theta) \geq I_A^{\tilde{b}}(\theta)$. Hence, raising the benchmark raises the amount of symmetric information the designer uses to induce investment on all agreement states.*

The economic intuition is simple. Under σ^b, the mass of anchors induced on each state θ maximizes agents' total perceived payoff from investment, holding fixed the total investment at $I^{\tilde{b}}(\theta)$, $I_A u(\tilde{b}(\theta), \theta) + \int_{I_A}^{I^{\tilde{b}}(\theta)} u(i, \theta) di$. A higher benchmark raises anchors' perceived payoff from investing, and so raises the mass of anchors induced.

5.4 Perfect Coordination

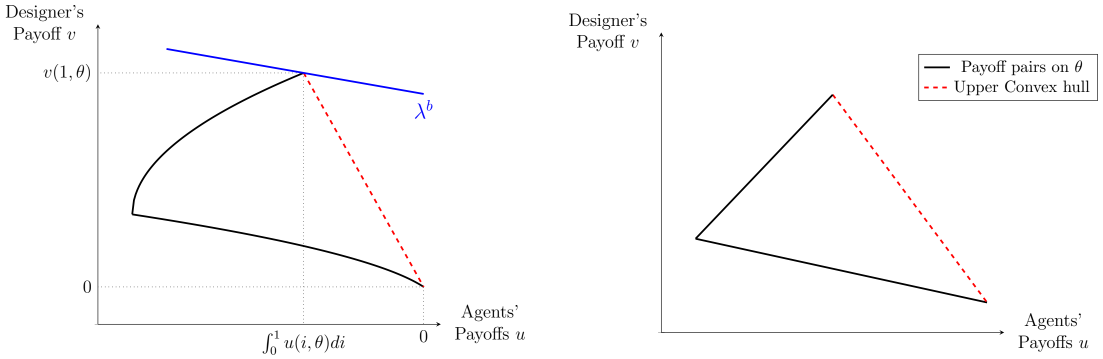
Say that an introspective outcome σ induces *perfect coordination* on state θ if for all $(I_A, I_N) \in \text{supp}(\sigma(\cdot|\theta))$, $I_A + I_N \in \{0, 1\}$. That is, on θ, either all agents invest (as some mix of anchors and non-anchors) or all agents do not invest, so agents' behaviours are ex-post symmetric. I will characterize when such an introspective outcome is optimal, and shed light on where the benchmark plays a role.

Intuitively, for the designer to prefer perfect coordination on a state, either one of two conditions should hold. First, the designer is able to raise total investment to one while raising agents' perceived payoffs from investing. This holds if the state to be a b-agreement state. Second, the designer benefits from randomizing between having all agents invest or not over always having a fraction of agents invest. I introduce next a weak condition on the designer's payoff to guarantee this holds.

Formally, say that the designer's payoff is *relatively convex* on state θ if

$$\forall I \in [0,1], \quad \int_I^1 u(i, \theta) di < 0 \Rightarrow \frac{\int_0^I u(i, \theta) di}{\int_0^1 u(i, \theta) di} v(1, \theta) \geq v(I, \theta)$$

Relative convexity is equivalent to stating that the upper right convex hull of payoff pairs $\{(\int_0^I u(i, \theta) di, v(I, \theta))\}_{I \in [0,1]}$ is linear. This clearly holds if the designer's payoff is convex in I, as in Section 2. Meanwhile, as seen in Figure 6, relative convexity can still hold if (a) the designer's payoff is concave but not too concave in I, or (b) the designer's payoff has a threshold form: $v(I, \theta) = 1$ if $I \geq \underline{I}(\theta)$, and $v(I, \theta) = 0$ if $I \geq \underline{I}(\theta)$. The latter often rises in the study of optimal information design in regime change games (e.g., Li et al. (2023), Inostroza and Pavan (2023)).



(a) $v(I,\theta)$ is concave in I

(b) $v(I,\theta)$ satisfies threshold form

Figure 6: In both figures, the designer's payoff (also) satisfies relative convexity on the state. For these, I further assume that on the state, agents' payoffs take on the form in Example 1

Proposition 8 below states the main implication of relative convexity: the designer prefers to perfectly coordinate investment on the state *regardless* of the benchmark.

Proposition 8. *Suppose the designer's payoff is relatively convex on θ. Then for all benchmarks b, the optimal introspective outcome σ^b induces perfect coordination on θ.*

The intuition follows from Figure 6. If θ is not a b-agreement state, then by Lemma 4, the amount of investment I induced on θ is either the largest or smallest maximizer of $v(I,\theta) + \lambda^b \int_0^I u(i,\theta)di$. That is, the pair $(\int_0^I u(i,\theta)di, v(I,\theta))$ induced is either the largest or smallest intersection between the upper right convex hull and the highest indifference curve (line) with slope λ^b. Under relative convexity, these extreme intersections always correspond to having all agents invest or not.

As all agents invest under the optimal introspective outcome σ^b on all b-agreement states, Proposition 8 implies the following result.

Corollary 1. *Suppose the designer's payoff is relatively convex on all non b-agreement states $\theta \notin \overline{\Theta}^b$. Then, the optimal introspective outcome σ^b induces perfect coordination on all states.*

Observe that raising the benchmark expands the set of agreement states. Thus, if the conditions of Corollary 1 holds at some benchmark b, then the designer perfectly coordinates investments not only at b, but for all higher benchmarks. Hence, raising the benchmark maintains (or raises) the designer's incentive to perfect coordinate.

Further observe that if relative convexity holds on all non **0**-agreement states, then Corollary 1 implies the designer prefers to perfectly coordinate regardless of the benchmark. This is notable as many environments have this property, e.g., those described before Proposition 8. It is also notable as restricted convexity is a weaker sufficient condition than many others provided in the literature for perfect coordination to be optimal under designer-preferred or adversarial selection.[27] Corollary 1 generalizes these results, while showing that relative convexity is sufficient for perfect coordination to be optimal under "intermediate" selection (benchmarks).

I now provide a sharper characterization of the optimal introspective outcome σ^b when it perfectly coordinates investments. Assume that states are labelled so that[28]

$$\theta > \theta' \Rightarrow \frac{-\int_0^1 u(i,\theta)di}{v(1,\theta)} \leq \frac{-\int_0^1 u(i,\theta')di}{v(1,\theta')}$$

$-\int_0^1 u(i,\theta)di/v(1,\theta)$ captures the "cost-to-benefit" ratio of inducing all agents to invest over not on a non b-agreement state. Thus, the smaller the state, the less efficient it is to induce investment on it. It follows that under σ^b, all agents invest on all b-agreement states and large enough non b-agreement states $\theta \geq \underline{\theta}^b$, where[29]

$$\underline{\theta}^b \equiv \inf\left\{\theta \in [0,1] : \int_{\overline{\Theta}^b} U(b(\tilde{\theta}), 1 - b(\tilde{\theta})|b, \tilde{\theta})dF(\tilde{\theta}) + \int_{[\theta,1]\backslash\overline{\Theta}^b} \int_0^1 u(i,\tilde{\theta})didF(\tilde{\theta}) \geq 0\right\} \tag{20}$$

Furthermore, by (17), σ^b induces exactly $b(\theta)$ anchors on all agreement states, and $\underline{I}(\theta)$ anchors on all large enough non b-agreement states $\theta \geq \overline{\theta}^b$, where

$$\overline{\theta}^b \equiv \inf\left\{\theta \in [\hat{\theta}^b, 1]\backslash\overline{\Theta}^b : \int_{\overline{\Theta}^b} b(\tilde{\theta})u(b(\tilde{\theta}),\tilde{\theta})dF(\tilde{\theta}) + \int_{[\theta,1]\backslash\overline{\Theta}^b} \underline{I}(\tilde{\theta})u(\underline{I}(\tilde{\theta}),\tilde{\theta})dF(\tilde{\theta}) \geq 0\right\}$$

I summarize these observations in Corollary 2 below.

Corollary 2. *Suppose the designer's payoff is relatively convex on all non b-agreement*

[27] For example, translated into my setting, Section 7.1 of Arieli and Babichenko (2019) shows that under designer-preferred selection, perfectly coordination is optimal if $v(I,\theta)$ is convex in I. Likewise, Theorem 2 of Morris et al. (2024) (also see Morris et al. (2022a)) show that under adversarial selection, perfectly coordination is optimal if for all $\theta \in \Theta$ and $I \in [0,1]$, $\int_I^1 u(i,\theta)di < 0 \Rightarrow Iv(1,\theta) \geq v(I,\theta)$. Both imply that relative convexity holds on all non $\mathbf{0}$-agreement states.

[28] When $v(1,\theta) = 0$, I let $-\int_0^1 u(i,\theta)di/v(1,\theta) = \infty$ if $\int_0^1 u(i,\theta)di < 0$, $-\int_0^1 u(i,\theta)di/v(1,\theta) = -\infty$ if $\int_0^1 u(i,\theta)di > 0$, and $\int_0^1 u(i,\theta)di/v(1,\theta) = 0$ if $\int_0^1 u(i,\theta)di = 0$.

[29] When perfect coordination is optimal and states are ordered by the ratio $-\int_0^1 u(i,\theta)di/v(1,\theta)$, $\underline{\theta}^b$ defined here coincides with the definition of $\underline{\theta}^b$ in (16).

states $\theta \notin \overline{\Theta}^b$. Then, the optimal introspective outcome σ^b is given by

$$\sigma^b(\cdot|\theta) = \begin{cases} \delta_{(b(\theta),1-b(\theta))}, & \theta \in \overline{\Theta}^b \\ \delta_{(\underline{I}(\theta),1-\underline{I}(\theta))}, & \theta \in [\overline{\theta}^b, 1]\backslash\overline{\Theta}^b \\ \delta_{(0,1)}, & \theta \in [\underline{\theta}^b, \overline{\theta}^b)\backslash\overline{\Theta}^b \\ \delta_{(0,0)}, & \theta \in [0, \underline{\theta}^b)\backslash\overline{\Theta}^b \end{cases} \tag{21}$$

Finally, I find that the mass of anchors under σ^b is now non-decreasing in the benchmark on *all* states. This strengthens the result of Proposition 7.

Corollary 3. *Suppose the designer's payoff is relatively convex on all non $\tilde{b}$-agreement states $\theta \notin \overline{\Theta}^{\tilde{b}}$. Then for all benchmarks $b \geq \tilde{b}$ and all $\theta \in \Theta$, $I_A^b(\theta) \geq I_A^{\tilde{b}}(\theta)$. Hence, raising the benchmark from $\tilde{b}$ raises the amount of symmetric information the designer uses to induce investment on all states.*

5.5 Welfare Implications

I now discuss the welfare impacts of raising the benchmark. To start, recall by Proposition 6 that doing so raises the aggregate investment across all states. Since the designer's payoff is increasing in aggregate investment, the following holds.

Corollary 4. *Take any two benchmarks $b, \tilde{b} \in \mathcal{B}$. If $b \geq \tilde{b}$, then $V^*(b) \geq V^*(\tilde{b})$. Hence, the designer benefits from a higher benchmark.*

The impact on agents' (ex-ante) payoffs is more subtle, and is the focus on this section. To simplify the exposition, I assume restricted convexity holds on all non **0**-agreement states, so the optimal introspective outcome σ^b is fully characterized by Corollary 2.

Combining Proposition 6 and Corollary 2, raising the benchmark from $\tilde{b}$ to $b \geq \tilde{b}$ only raises the set of states on which (all) agents invest on under the optimal introspective outcome σ^b. In fact, the additional set of states is $([\underline{\theta}^b, \underline{\theta}^{\tilde{b}}] \cup \overline{\Theta}^b)\backslash\overline{\Theta}^{\tilde{b}}$. Thus, the change in agents' ex-ante payoffs is

$$\int_{([\underline{\theta}^b, \underline{\theta}^{\tilde{b}}] \cup \overline{\Theta}^b)\backslash\overline{\Theta}^{\tilde{b}}} u(1,\theta) dF(\theta) \tag{22}$$

By (22), if $u(1,\theta) \geq 0$ holds for all states in $([\underline{\theta}^b, \underline{\theta}^{\tilde{b}}] \cup \overline{\Theta}^b)\backslash\overline{\Theta}^{\tilde{b}}$, then agents benefit from any increase in the benchmark. Likewise, if $u(1,\theta) \leq 0$ holds for all such states, as in the example of Section 2, then raising the benchmark harms agents.

In general, neither of the preceding cases holds. The next result speaks to these cases. It shows that raising the benchmark on specific states can (still) lead to monotone changes in agents' payoffs. To state the result, call a state θ *costly for investment at b* if $\theta \notin [\underline{\theta}^b, 1] \cup \overline{\Theta}^b$. That is, state θ is both not a b-agreement state and has a high cost-to-benefit ratio $-\int_0^1 u(i,\theta)di/v(1,\theta)$, and so "costly" to induce investment on.

Proposition 9. *Suppose the designer's payoff is relatively convex on all non b-agreement states $\theta \notin \overline{\Theta}^b$. Take any two benchmarks $b, \tilde{b} \in \mathcal{B}$ with $b \geq \tilde{b}$.*

1. *If $u(b(\theta),\theta) > u(\tilde{b}(\theta),\theta)$ holds only on states which are costly for investment under $\tilde{b}$, then agents are weakly better off under σ^b than $\sigma^{\tilde{b}}$. If, in addition,*

$$\int_{([\underline{\theta}^b,\underline{\theta}^{\tilde{b}}]\cup\overline{\Theta}^b)\setminus\overline{\Theta}^{\tilde{b}}} (u(1,\theta) - U(I^b_A(\theta), 1 - I^b_A(\theta)|b,\theta))dF(\theta) > 0 \tag{23}$$

   holds, then agents are strictly better off under σ^b than $\sigma^{\tilde{b}}$.

2. *If $u(\tilde{b}(\theta),\theta) = u(1,\theta)$ holds on all states costly for investment under $\tilde{b}$, then agents are weakly worse off under σ^b than $\sigma^{\tilde{b}}$. If, in addition, agents invest on strictly more states, so $([\underline{\theta}^b, \underline{\theta}^{\tilde{b}}] \cup \overline{\Theta}^b)\setminus\overline{\Theta}^{\tilde{b}}$ is of strictly positive measure, then agents are strictly worse off under σ^b than $\sigma^{\tilde{b}}$.*

Part 1 says that raising the benchmark (from $\tilde{b}$) only on states costly for investment never leaves agents worse off. The intuition is as follows. First, I show that the total change in agents' total perceived payoffs from investment under the optimal introspective outcome must be non-negative. Then, I observe that this change comes from two sources: the change on states where the designer continues to induce all agents to invest on, and the change on the additional set of states where agents now invest on. The first source is zero, because those states are the states not costly for investment under $\tilde{b}$. The second source, which must then be non-negative, is

$$\int_{([\underline{\theta}^b,\underline{\theta}^{\tilde{b}}]\cup\overline{\Theta}^b)\setminus\overline{\Theta}^{\tilde{b}}} U(I^b_A(\theta), 1 - I^b_A(\theta))|b,\theta)dF(\theta) \tag{24}$$

where $U(I^b_A(\theta), 1 - I^b_A(\theta))|b,\theta)$ is agents' perceived payoffs from investment under the mass of anchors and non-anchors induced on state θ under the optimal introspective outcome σ^b. Since $u(1,\theta) \geq U(I_A^b(\theta), 1 - I_A^b(\theta))|b, \theta)$, the change in agents' ex-ante payoffs in (22) is weakly greater than (24), and strictly so if (23) holds.

Part 2 says that if raising the benchmark from $\tilde{b}$ on states costly for investment no longer raises' agents' perceived payoffs from investment, so $u(\tilde{b}(\theta), \theta) = u(1, \theta)$ holds, then further raising the benchmark never leaves agents better off. This is because all states costly for investment cannot be $\tilde{b}$-agreement states, which means $u(1,\theta) < 0$ holds. As such, the change agents' ex-ante payoffs in (22) must be weakly negative, and strictly so if they invest on strictly more states.

Parts 1 and 2 have a sequential interpretation: one can first raise the benchmark as in Part 1 to raise agents' payoffs, and then decreases it as in Part 2 to lower payoffs. This suggests that moderate increases in the benchmark benefit agents, but excessive ones do not. I illustrate this with an example.

Example 5. *Suppose there are three states $\theta \in \{L, M, H\}$, where $Pr(\theta = L) = 1/4$, $Pr(\theta = M) = 1/4$ and $Pr(\theta = H) = 1/2$. Furthermore, $v(I, \theta) = I$ and*

$$u(I, H) = 1/32, \quad u(I, M) = \begin{cases} -10, & I < 1/2 \\ \frac{1+I}{2}, & I \geq 1/2 \end{cases}, \quad u(I, L) = -1$$

First, consider the lowest benchmark $b = \mathbf{0}$. Here, the optimal introspective outcome σ^b satisfies $\sigma^b(\cdot|H) = \delta_{(0,1)}$, $\sigma^b(\cdot|M) = \delta_{(0,0)}$ and $\sigma^b(\cdot|L) = 1/64 \times \delta_{(0,1)} + 63/64 \times \delta_{(0,0)}$. That is, the designer always induces investment on state H, never on state M, and sometimes on state L. Thus, agents' ex-ante payoffs under it is zero.

Now raise the benchmark to $b \geq \mathbf{0}$ where $b(H) = b(L) = 0$ but $b(M) = 1/2$. Here, the benchmark only rises on state M, which is costly for investment under $\mathbf{0}$, so Part 1 of Proposition 9 applies. The optimal introspective outcome is $\sigma^b(\cdot|H) = \delta_{(0,1)}$, $\sigma^b(\cdot|M) = \delta_{(1/2,1/2)}$ and $\sigma^b(\cdot|L) = 27/64 \times \delta_{(0,1)} + 37/64 \times \delta_{(0,0)}$. That is, the designer now induces all agents to invest on state M, and more frequently on state L. This increases agents' ex-ante payoffs to $1/32 \times 1/4 + 1 \times 1/4 + -1 \times 1/2 \times 27/64 = 3/64 > 0$.

Finally, raise the benchmark to $\mathbf{1}$. Here, $u(b(\theta), \theta) = u(1, \theta) = -1$ on state $\theta = L$, the only state costly for investment under b, so Part 2 of Proposition 9 applies. The optimal introspective outcome is $\sigma^b(\cdot|H) = \delta_{(0,1)}$, $\sigma^b(\cdot|M) = \delta_{(1/2,1/2)}$ and $\sigma^b(\cdot|L) = 33/64 \times \delta_{(0,1)} + 31/64 \times \delta_{(0,0)}$. That is, the designer induces all agents to invest more frequently on state L. This decreases agents' ex-ante payoffs to 0.

5.6 Non-Threshold Games

In the supplementary appendix, I extend several key results from the previous sections to the case where Assumption 1 does not hold. I briefly discuss these next.

First, given benchmark b, I show that there exists an optimal introspective outcome which has all agents invest on all b-agreement states. Furthermore, their total perceived payoffs from investing on such states is positive. Thus, raising the benchmark to b weakly raises aggregate investment on at least all b-agreement states, and generates additional incentives for agents to invest, which the designer can "spend" to raise investment on other states. This is the general counterpart of Proposition 6.

Second, I show that under any optimal introspective outcome, if $I_A < 1$ anchors are drawn on a b-agreement state θ, and $I'_A < 1$ anchors are drawn on a non b-agreement state θ', then $u(I_A, \theta) \geq 0$ and

$$\frac{u(I_A, \theta)}{u(b(\theta), \theta)} \leq \frac{u(I'_A, \theta')}{u(b(\theta'), \theta')}$$

This condition imposes a lower bound on the mass of anchors induced on both states, which itself depends on the benchmark. In particular, fixing I_A and I'_A a large enough increase in $b(\theta)$ and $b(\theta')$ will cause the inequality to be violated. Thus, provided the designer continues to induce investment on states θ and θ', raising the benchmark by enough increases the mass of anchors induced on at least one of the two states. This is the general counterpart of Propositions 7 and Corollary 3.

Finally, say that the designer's payoff satisfies *weak convexity* if for all $I \in [0, 1]$, $v(I, \theta) \leq Iv(1, \theta)$, which is a stronger condition than relative convexity. I show that if the designer's payoff satisfies weak convexity on all non b-agreement states, then there exists an optimal introspective outcome that perfectly coordinates investments across all states. This is the general counterpart of Corollary 1.

6 On Public Information Design

This section explores the consequences of restrcting the designer to public information structures, where all agents always observe the same signal.[30] There are many environments where this is relevant. For example, an entrepreneur may raise invest-

[30] Formally, $\mathcal{S}$ is a public information structure if for all $\theta \in \Theta$, $\text{supp}(\pi(\cdot|\theta)) \subseteq \{\delta_s\}_{s \in S}$.

ment via an IPO, where all information about the project must be publicly available. Likewise, banking stress test results must often be public. Finally, agents may communicate amongst each other prior to investment.

To simplify the analysis, I impose Assumption 2 and the following three assumptions. Together, these define a generalization of Section 2's setting.

Assumption 4. *There exists a state* $\overline{\theta} \in (0,1)$ *such that*

1. *For all* $\theta \geq \overline{\theta}$, $u(0,\theta) \geq 0$ *and* $u(I,\theta)$ *is strictly increasing in* I.

2. *For all* $\theta < \overline{\theta}$, $u(1,\theta) = u(0,\theta) \equiv \underline{u}(\theta) < 0$. *Additionally, for all* $\theta, \tilde{\theta} < \overline{\theta}$, $\theta > \tilde{\theta}$ *implies* $-\underline{u}(\theta)/v(1,\theta) \leq -\underline{u}(\tilde{\theta})/v(1,\tilde{\theta})$.

Assumption 5. *For all* $\theta \in \Theta$, $v(I,\theta)$ *is strictly increasing in* I. *Furthermore,* $\int_{\Theta} u(1,\theta)dF(\theta) \leq 0$.

Assumption 6. *The designer's payoff satisfies relative convexity on all* $\theta \notin \overline{\Theta}^{\mathbf{0}}$.

Assumption 4 is a special case of a threshold game where agents have a dominant action on each state. Assumption 5 implies the designer strictly prefers to have all agents invest on all states, but may not be able to achieve it as investing is on average weakly dominated for agents. Finally, Assumption 6 implies the designer prefers to perfectly coordinate investments. Thus, the optimal introspective outcome σ^b, which solves the unconstrained information design problem, is characterized by Corollary 2.

The public information design problem is simpler than the general problem. This is as under public information, whether (all) agents invest depends solely on whether they invest against the benchmark. In turn, the designer's problem reduces to persuading a representative agent to invest, where the agent has payoffs of $u(b(\theta),\theta)$ and 0 from investing and not investing on state θ, and the designer has a payoff of $v(1,\theta)$ and 0 when the agent invests and not invests respectively. I show in Appendix C that the optimal public information structure then has a simple form: all agents invest if and only if the state θ exceeds a threshold $\tilde{\theta}^b$. Furthermore, the threshold is non-increasing in the benchmark b, so the designer benefits from a higher benchmark.

Given the above, it is important to understand when public information is optimal. Note that a b-obedient introspective outcome can be implemented by public information if and only if perfectly coordinates investments and never draws non-anchors. By Corollary 2, this holds under the optimal introspective outcome σ^b for $b = \mathbf{1}$, so public information is optimal at the highest benchmark. Meanwhile, the next result shows that for most other benchmarks, public information is strictly suboptimal.

Proposition 10. *Suppose there exists a measurable set of states $\tilde{\Theta} \subseteq [0, \overline{\theta}]$ with (i) $F(\tilde{\Theta}) > 0$, and (ii) $b(\theta) < 1$ for all $\theta \in \tilde{\Theta}$. Then, every public information structure is strictly suboptimal for the designer.*

The intuition is similar to Section 2. When the benchmark is strictly lower than one on a non-trivial set of agreement states, so (i) and (ii) hold, then the designer cannot induce all agents to invest on all states under any public information structure. Meanwhile, under any such information structure, agents' payoffs from investing at higher rounds of introspection is strictly positive. The designer can leverage this slack via private information to induce agents to invest on a strictly larger set of states.

Let $V^{\text{Pub}}(b)$ denote the designer's payoff under an optimal public information structure given benchmark b. The next result says that the loss from being restricted to public information, $V^*(b) - V^{\text{Pub}}(b)$, is smaller under a higher benchmark.

Proposition 11. *Take any two benchmarks $b, \tilde{b} \in \mathcal{B}$. If $b \geq \tilde{b}$, then $V^*(b) - V^{Pub}(b) \leq V^*(\tilde{b}) - V^{Pub}(\tilde{b})$.*

The intuition is as follows. Under a low benchmark, the designer benefits from substituting some anchors for non-anchors via using private information. Hence, an optimal public information structure induces more anchors than under the optimal introspective outcome. Thus, raising the benchmark, which raises only anchors' incentives to invest, allows the designer to raise agents' investment by far more on other states when the designer is constrained to public information than when she is not.

Given benchmarks $b \geq \tilde{b}$, Proposition 11 also implies $V^*(b) - V^*(\tilde{b}) \leq V^{\text{Pub}}(b) - V^{\text{Pub}}(\tilde{b})$. That is, the designer's marginal gain from a higher benchmark is higher when restricted to public information. Going back to the entrepreneur example, this suggests that benchmarks have a greater impact on investment outcomes when funds are raised via an IPO than when privately engaging with investors.

7 Conclusion

In this paper, I develop a framework for studying the impact of benchmarks, which affect agents' reasoning about others' behaviours, on optimal information disclosure in binary action supermodular games. I show that the approach is general enough to nest prior analysis, and tractable by characterizing the possible outcomes that arise. For a large class of games, I construct an optimal information structure, and discuss how it varies in the benchmark and its consequences on the outcome. Finally, I demonstrate the interaction between the benchmark and informational constraints.

The key assumption of my model is that benchmarks are fixed. A natural extension is to allow the designer to influence agents' benchmarks. For instance, an entrepreneur can choose to raise funding from investors with different benchmarks. In the supplementary appendix, I allow the designer to choose the agent's benchmark under a given signal observed, subject to an upper bound on the aggregate benchmark induced across agents. There, I show how the designer's problem can also be reformulated into optimizing over implementable generalized introspective outcomes, characterized by obedience constraints. I also find that whether the designer optimally induces different benchmarks depends critically on the marginal returns from investment. For example, in the setting of Section 6, the designer optimally gives all anchors the same benchmark when agents' payoffs from investment $u(I, \theta)$ are concave in I and submodular in (I, θ).

Other possible extensions involve giving less control over the benchmark to the designer. For example, the benchmark may vary directly in the designer's choice of information structure. For instance, public disclosure may act as a focal point for investors (Morris and Shin, 2002), leading to high benchmarks. In the spirit of global games (Carlsson and van Damme, 1993), agents' benchmarks can also be determined by an exogenous signal that is informative about the state. For example, by researching into the entrepreneur, an investor both develops a benchmark and learns about the entrepreneur's project quality. I leave these avenues for future research.

References

Akerlof, Robert and Richard Holden, "Movers and Shakers," *The Quarterly Journal of Economics*, 2016, *131* (4), 1849–1874.

_ **and** _ , "Capital Assembly," *The Journal of Law, Economics, and Organization*, 07 2019, *35* (3), 489–512.

_ **and** _ , "Coordinating Supply Chains," *Working paper*, 2023.

_ , _ , and Luis Rayo, "Network Externalities and Market Dominance," *Management Science*, 2023.

Alesina, Alberto and Paola Giuliano, "Culture and Institutions," *Journal of Economic Literature*, December 2015, *53* (4), 898–944.

Arieli, Itai and Yakov Babichenko, "Private Bayesian persuasion," *Journal of Economic Theory*, 2019, *182*, 185–217.

Arrow, Kenneth J, *The Theory of Discrimination*, Princeton University Press, 1973.

Aumann, Robert J., "Subjectivity and correlation in randomized strategies," *Journal of Mathematical Economics*, 1974, *1* (1), 67–96.

Bergemann, Dirk and Stephen Morris, "Bayes correlated equilibrium and the comparison of information structures in games," *Theoretical Economics*, 2016, *11* (2), 487–522.

_ and _ , "Information Design: A Unified Perspective," *Journal of Economic Literature*, March 2019, *57* (1), 44–95.

Brooks, Alison Wood, Laura Huang, Sarah Wood Kearney, and Fiona E. Murray, "Investors prefer entrepreneurial ventures pitched by attractive men," *Proceedings of the National Academy of Sciences*, 2014, *111* (12), 4427–4431.

Candogan, Ozan and Kimon Drakopoulos, "Optimal Signaling of Content Accuracy: Engagement vs. Misinformation," *Operations Research*, 2020, *68* (2). Published Online: 2 Mar 2020.

Carlsson, Hans and Eric van Damme, "Global Games and Equilibrium Selection," *Econometrica*, 1993, *61* (5), 989–1018.

Carroll, Gabriel, "Informationally robust trade and limits to contagion," *Journal of Economic Theory*, 2016, *166*, 334–361.

Chassang, Sylvain, "Building Routines: Learning, Cooperation, and the Dynamics of Incomplete Relational Contracts," *American Economic Review*, March 2010, *100* (1), 448–65.

Colombo, Massimo G., Chiara Franzoni, and Claudia Rossi-Lamastra, "Internal Social Capital and the Attraction of Early Contributions in Crowdfunding," *Entrepreneurship Theory and Practice*, 2015, *39* (1), 75–100.

Crawford, Vincent P., Miguel A. Costa-Gomes, and Nagore Iriberri, "Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications," *Journal of Economic Literature*, March 2013, *51* (1), 5–62.

Crosetto, Paolo and Tobias Regner, "It's Never Too Late: Funding Dynamics and Self-Pledges in Reward-Based Crowdfunding," *Research Policy*, 2018, *47* (8), 1463–1477.

Crémer, Jacques, " Corporate Culture and Shared Knowledge *," *Industrial and Corporate Change*, 09 1993, *2* (3), 351–386.

Ewens, Michael and Richard R. Townsend, "Are early stage investors biased against women?," *Journal of Financial Economics*, 2020, *135* (3), 653–677.

Frankel, David M., Stephen Morris, and Ady Pauzner, "Equilibrium selection in global games with strategic complementarities," *Journal of Economic Theory*, 2003, *108* (1), 1–44.

Gibbons, Robert and Rebecca Henderson, *17. What Do Managers Do?: Exploring Persistent Performance Differences among Seemingly Similar Enterprises*, Princeton: Princeton University Press,

Goldstein, Itay and Chong Huang, "Bayesian Persuasion in Coordination Games," *American Economic Review*, May 2016, *106* (5), 592–96.

Halac, Marina, Elliot Lipnowski, and Daniel Rappoport, "Rank Uncertainty in Organizations," *American Economic Review*, March 2021, *111* (3), 757–786.

_ , _ , and _ , "Addressing Strategic Uncertainty with Incentives and Information," *AEA Papers and Proceedings*, May 2022, *112*, 431–37.

Hebert, Camille, "Gender Stereotypes and Entrepreneur Financing," *Working paper*, 2023.

Hoshino, Tetsuya, "Multi-Agent Persuasion: Leveraging Strategic Uncertainty," *International Economic Review*, 2022, *63* (2), 755–776.

Inostroza, Nicolas and Alessandro Pavan, "Adversarial Coordination and Public Information Design," *Working paper*, 2023.

Kajii, Atsushi and Stephen Morris, "The Robustness of Equilibria to Incomplete Information," *Econometrica*, 1997, *65* (6), 1283–1309.

Kets, Willemien, "Organizational Design: Culture and Incentives," *Working paper*, 2021.

_ and Alvaro Sandroni, "A Theory of Strategic Uncertainty and Cultural Diversity," *The Review of Economic Studies*, 08 2020, *88* (1), 287–333.

_ , Wouter Kager, and Alvaro Sandroni, "The value of a coordination game," *Journal of Economic Theory*, 2022, *201*, 105419.

Li, Fei, Yangbo Song, and Mofei Zhao, "Global manipulation by local obfuscation," *Journal of Economic Theory*, 2023, *207*, 105575.

Lipnowski, Elliot, Doron Ravid, and Denis Shishkin, "Perfect Bayesian Persuasion," 2024.

Mathevet, Laurent, Jacopo Perego, and Ina Taneva, "On Information Design in Games," *Journal of Political Economy*, 2020, *128* (4), 1370–1404.

Miishe, Addy, "How women African entrepreneurs can overcome the "beauty pageant problem," https://oecd-development-matters.org/2022/06/21/how-women-african-entrepreneurs-can-overcome-the-beauty-pageant-problem/ 2022. Accessed: 2024-07-07.

Milgrom, Paul and Ilya Segal, "Envelope Theorems for Arbitrary Choice Sets," *Econometrica*, 2002, *70* (2), 583–601.

_ **and John Roberts**, "Rationalizability, Learning, and Equilibrium in Games with Strategic Complementarities," *Econometrica*, 1990, *58* (6), 1255–1277.

Moriya, Fumitoshi and Takuro Yamashita, "Asymmetric-information allocation to avoid coordination failure," *Journal of Economics & Management Strategy*, 2020, *29* (1), 173–186.

Morris, Stephen and Hyun Song Shin, "Social Value of Public Information," *American Economic Review*, December 2002, *92* (5), 1521–1534.

_ **and Hyun Song Shin**, *Global games: Theory and applications*, United Kingdom: Cambridge University Press, January

_ **and Takashi Ui**, "Generalized Potentials and Robust Sets of Equilibria," *Journal of Economic Theory*, 2005, *124* (1), 45–78.

_ **, Daisuke Oyama, and Satoru Takahashi**, "Implementation via Information Design using Global Games," *Working paper*, 2022.

_ **,** _ **, and** _ , "On the Joint Design of Information and Transfers," *Working paper*, 2022.

_ **,** _ **, and** _ , "Implementation via Information Design in Binary-Action Supermodular Games," *Econometrica, Forthcoming*, 2024.

Oyama, Daisuke and Satoru Takahashi, "Generalized Belief Operator and Robustness in Binary-Action Supermodular Games," *Econometrica*, 2020, *88* (2), 693–726.

Phelps, Edmund S, "The Statistical Theory of Racism and Sexism," *American Economic Review*, 1972, *62* (4), 659–661.

Rubinstein, Ariel, "The Electronic Mail Game: Strategic Behavior under 'Almost

Common Knowledge',” *American Economic Review*, 1989, *79* (3), 385–391.

Segal, Ilya, “Contracting with Externalities,” *The Quarterly Journal of Economics*, 1999, *114* (2), 337–388.

_ , “Coordination and discrimination in contracting with externalities: divide and conquer?,” *Journal of Economic Theory*, 2003, *113* (2), 147–181.

Sun, Yeneng, “The exact law of large numbers via Fubini extension and characterization of insurable risks,” *Journal of Economic Theory*, 2006, *126* (1), 31–69.

Taneva, Ina, “Information Design,” *American Economic Journal: Microeconomics*, November 2019, *11* (4), 151–85.

_ **and Laurent Mathevet**, “Organized Information Transmission,” *Working paper*, 2023.

Winter, Eyal, “Incentives and Discrimination,” *American Economic Review*, June 2004, *94* (3), 764–773.

Ziegler, Gabriel, “Adversarial Bilateral Information Design,” *Working Paper*, 2020.

Appendix A: Proofs for Implementation

This section contains the proofs of Theorems 1 and 2. Throughout, I let π_S denote the marginal distribution over signals $s \in S$ induced by an information structure $\mathcal{S}$.

A1: Proof of Theorem 1

Take a b-implementable introspective outcome σ and any information structure $\mathcal{S}$ that implements it. I will show that it satisfies b-obedience.

First, let $S^1 \equiv \{s \in S : \alpha^{\mathcal{S},b,1}(s) = 1\}$ denote the set of signals on which agents are anchors. Then,

$$\begin{aligned}
\int_\Theta \int_{\mathcal{I}} I_A u(b(\theta),\theta) d\sigma(I_A, I_N|\theta) dF(\theta) &= \int_\Theta \int_{\Delta(S)} I_A^{\mathcal{S},b}(\mu) u(b(\theta),\theta) d\pi(\mu|\theta) dF(\theta) \\
&= \int_\Theta \int_{\Delta(S)} \int_{S^1} u(b(\theta),\theta) d\pi(s,\mu,\theta) \\
&= \int_{S^1} \underbrace{\int_{\Delta(S)\times\Theta} u(b(\theta),\theta) d\pi(\mu,\theta|s)}_{\geq 0 \text{ by equation (1)}} d\pi_S(s) \geq 0
\end{aligned}$$

where the first equality holds as $\sigma(\cdot|\theta)$ is the pushforward of $\pi(\cdot|\theta)$ through $I^{\mathcal{S},b}(\cdot)$, and the second equality uses $I_A^{\mathcal{S},b}(\mu)u(b(\theta),\theta) = \int_{S^1} u(b(\theta),\theta)d\mu(s)$. Likewise,

$$\begin{aligned}\int_\Theta \int_{\mathcal{I}} I_A u(I_A,\theta) d\sigma(I_A, I_N|\theta) dF(\theta) &= \int_\Theta \int_{\Delta(S)} I_A^{\mathcal{S},b}(\mu) u(I_A^{\mathcal{S},b}(\mu),\theta) d\pi(\mu|\theta) dF(\theta) \\ &= \int_\Theta \int_{\Delta(S)} \int_{S^1} u(I_A^{\mathcal{S},b}(\mu),\theta) d\pi(s,\mu,\theta) \\ &= \int_{S^1} \underbrace{\int_{\Delta(S)\times\Theta} u(I_A^{\mathcal{S},b}(\mu),\theta) d\pi(\mu,\theta|s)}_{\geq 0 \text{ by equation (2)}} d\pi_S(s) \geq 0\end{aligned}$$

Hence, anchor obedience holds.

Next, for each $k > 1$, let $S^k \equiv \{s \in S : \alpha^{\mathcal{S},b,k}(s) = 1 > 0 = \alpha^{\mathcal{S},b,k-1}(s)\}$ denote the set of signals on which an agent switches to invest at round k of introspection. By aggregating such agents' payoffs from switching, notice that

$$\sum_{k=2}^\infty \int_{S^k} u(I(\mu|\alpha^{\mathcal{S},b,k-1}),\theta) d\mu(s) \leq \sum_{k=2}^\infty \int_{I(\mu|\alpha^{\mathcal{S},b,k-1})}^{I(\mu|\alpha^{\mathcal{S},b,k})} u(i,\theta) di = \int_{I_A^{\mathcal{S},b}(\mu)}^{I_A^{\mathcal{S},b}(\mu)+I_N^{\mathcal{S},b}(\mu)} u(i,\theta) di =$$

where the inequality holds as $u(I,\theta)$ is non-decreasing in I and $I(\mu|\alpha^{\mathcal{S},b,k}) - I(\mu|\alpha^{\mathcal{S},b,k-1}) = \int_{S^k} d\mu(s)$, and the equality holds as $I(\mu|\alpha^{\mathcal{S},b,1}) = I_A^{\mathcal{S},b}(\mu)$ while $\lim_{k\to\infty} I(\mu|\alpha^{\mathcal{S},b,k}) = I_A^{\mathcal{S},b}(\mu) + I_N^{\mathcal{S},b}(\mu)$. Therefore,

$$\begin{aligned}\int_\Theta \int_{\mathcal{I}} \left(\int_{I_A}^{I_A+I_N} u(i,\theta) di \right) d\sigma(I_A, I_N|\theta) dF(\theta) &= \int_\Theta \int_{\Delta(S)} \int_{I_A^{\mathcal{S},b}(\mu)}^{I_A^{\mathcal{S},b}(\mu)+I_N^{\mathcal{S},b}(\mu)} u(i,\theta) di d\pi(\mu|\theta) dF(\theta) \\ &\geq \int_\Theta \int_{\Delta(S)} \sum_{k=2}^\infty \int_{S^k} u(I(\mu|\alpha^{\mathcal{S},b,k-1}),\theta) d\mu(s) d\pi(\mu|\theta) dF(\theta) \\ &= \sum_{k=2}^\infty \left(\int_\Theta \int_{\Delta(S)} \int_{S^k} u(I(\mu|\alpha^{\mathcal{S},b,k-1}),\theta) d\pi(s,\mu,\theta) \right) \\ &= \sum_{k=2}^\infty \int_{S^k} \underbrace{\left(\int_{\Delta(S)\times\Theta} u(I(\mu|\alpha^{\mathcal{S},b,k-1}),\theta) d\pi(\mu,\theta|s) \right)}_{\geq 0 \text{ by equation (2)}} d\pi_S(s) \geq 0\end{aligned}$$

where the order of summation and integration in the second equality can be switched by the Fubini-Tonelli Theorem. Therefore, non-anchor obedience holds.

It remains to show downwards obedience. First, let $S^{NA} \equiv \cup_{k\geq 2} S^k$. Then,

$$\int_\Theta \int_{\mathcal{I}} I_N u(b(\theta),\theta) d\sigma(I_A, I_N|\theta) dF(\theta) = \int_\Theta \int_{\Delta(S)} I_N^{\mathcal{S},b}(\mu) u(b(\theta),\theta) d\pi(\mu|\theta) dF(\theta)$$

$$\begin{aligned}
&= \int_\Theta \int_{\Delta(S)} \int_{S^{NA}} u(b(\theta),\theta) d\mu(s) d\pi(\mu|\theta) dF(\theta) \\
&= \int_\Theta \int_{\Delta(S)} \int_{S^{NA}} u(b(\theta),\theta) d\pi(s,\mu,\theta) \\
&= \int_{S^{NA}} \underbrace{\int_{\Delta(S)\times\Theta} u(b(\theta),\theta) d\pi(\mu,\theta|s)}_{<0 \text{ by equation (1)}} d\pi_S(s) \leq 0
\end{aligned}$$

where the second equality holds as $I_N^{\mathcal{S},b}(\mu) u(b(\theta),\theta) = \int_{S^{NA}} u(b(\theta),\theta) d\mu(s)$, and the last inequality is strict if $\pi(S^{\text{NA}}) > 0$, i.e., $\int_\Theta \int_\mathcal{I} I_N d\sigma(I_A, I_N|\theta) dF(\theta) > 0$. Hence, (9) holds.

Next, applying a similar logic to the above but replacing S^{NA} with $S^{NI} \equiv \{s : \alpha^{\mathcal{S},b}(s) = 0\}$, $I_N^{\mathcal{S},b}(\mu)$ with $1 - I_A^{\mathcal{S},b}(\mu) - I_N^{\mathcal{S},b}(\mu)$, and I_N with $1 - I_A - I_N$ whenever they arise, one finds that (10) holds.

Finally, to see (11) holds, take any signal $s \in S$ under which $\alpha^{\mathcal{S},b}(s) = 0$. Then, $\alpha^{\mathcal{S},b,k}(s) = 0$ for all $k \geq 1$. Furthermore, as $u(I,\theta)$ is non-decreasing in I for each θ, and since $I_A^k(\mu) + I_N^k(\mu) \equiv \int_S \alpha^{\mathcal{S},b,k}(s) d\mu(s)$ is non-decreasing in k for each μ, $(u(I_A^k(\mu) + I_N^k(\mu), \theta))_{k \geq 2}$ is a sequence of measurable functions on $\Delta(S) \times \Theta$ that converges monotonically point-wise to $u(I_A^{\mathcal{S},b}(\mu) + I_N^{\mathcal{S},b}(\mu), \theta)$, and is bounded above and below by the integrable functions $u(1,\theta)$ and $u(0,\theta)$ respectively. Hence, by the Dominated Convergence Theorem,

$$\int_{\Delta(S)\times\Theta} u(I_A^{\mathcal{S},b}(\mu) + I_N^{\mathcal{S},b}(\mu), \theta) d\pi(\mu,\theta|s) = \lim_{k\to\infty} \underbrace{\int_{\Delta(S)\times\Theta} u(I_A^k(\mu) + I_N^k(\mu), \theta) d\pi(\mu,\theta|s)}_{\leq 0 \text{ as } \alpha^{\mathcal{S},b,k}(s)=0 \text{ for all } k} \leq 0$$

Therefore,

$$\begin{aligned}
\int_\Theta \int_\mathcal{I} (1 - I_A - I_N) u(I_A + I_N, \theta) d\sigma(I_A, I_N|\theta) dF(\theta) &= \int_\Theta \int_{\Delta(S)} \int_{S^3} u(I_A^{\mathcal{S},b}(\mu) + I_N^{\mathcal{S},b}(\mu), \theta) d\mu(s) d\pi(\mu|\theta) dF(\theta) \\
&= \int_{S^{NI}} \left(\int_{\Delta(S)\times\Theta} u(I_A^{\mathcal{S},b}(\mu) + I_N^{\mathcal{S},b}(\mu), \theta) d\pi(\mu,\theta|s) \right) d\pi_S(s) \leq 0
\end{aligned}$$

Hence, downwards obedience holds. □

A2: Proof of Theorem 2

Take any b-obedient introspective outcome σ. The goal is to prove that σ is approximately implementable. Section A2.1 proves that this is true under certain restrictions on σ. Section A2.2 then extends the proof to the general case.

Before proceeding, observe that if $\int_\Theta I_N d\tilde{\sigma}(I_A, I_N|\theta)dF(\theta) = 0$ holds, then there exists an introspective outcome $\tilde{\sigma}$ equivalent to σ, i.e., $\|\sigma - \tilde{\sigma}\| = 0$, in which for all $(I_A, I_N) \in \text{supp}(\tilde{\sigma}(\cdot|\theta))$, $I_N = 0$. That is, it suffices to assume that under σ, $\int_\Theta I_N d\tilde{\sigma}(I_A, I_N|\theta)dF(\theta) = 0$ implies non-anchors are never drawn across states. Likewise, I assume that if $\int_\Theta I_N d\tilde{\sigma}(I_A, I_N|\theta)dF(\theta) = 0$, then non-investors are never drawn on all states. I will also make use of the following auxilliary result, which will be used to verify when an information structure has a b-MIE.

Lemma 5. *An information structure $\mathcal{S}$ has a b-MIE if and only if for all $s \in S$, if $\alpha^{\mathcal{S},b,1}(s) = 1$, then $\alpha^{\mathcal{S},b,2}(s) = 1$.*

Proof. The only if part is immediate. To prove the if part, it suffices to prove that for all $k \geq 2$, $\alpha^{\mathcal{S},b,k-1}(s) = 1$ implies $\alpha^{\mathcal{S},b,k}(s) = 1$. I will do so via induction on k. The base case for $k = 2$ holds by assumption. Hence, suppose the induction hypothesis holds at $k-1$ for some $k > 2$. Take any $s \in S$ in which $\alpha^{\mathcal{S},b,k-1}(s) = 1$. Then, since $\alpha^{\mathcal{S},b,k-1}(s) \geq \alpha^{\mathcal{S},b,k-2}(s) = 1$ by the induction hypothesis,

$$\int_{\Delta(S)\times\Theta} u(I(\mu|\alpha^{\mathcal{S},b,k-1}), \theta)d\pi(\mu, \theta|s) \geq \int_{\Delta(S)\times\Theta} u(I(\mu|\alpha^{\mathcal{S},b,k-2}), \theta)d\pi(\mu, \theta|s) \geq 0$$

and so $\alpha^{\mathcal{S},b,k}(s) = 1$. □

A2.1. Strict version of Theorem 2

I first prove Theorem 2 holds when σ also (i) satisfies anchor obedience strictly, (ii) satisfies non-anchor obedience strictly (if $\int_\Theta I_N d\tilde{\sigma}(I_A, I_N|\theta)dF(\theta) > 0$), (iii) satisfies downwards obedience in (11) strictly (if $\int_\Theta (1 - I_A - I_N) d\tilde{\sigma}(I_A, I_N|\theta)dF(\theta) > 0$) and (iv) for all $\theta \in \overline{\Theta}$, $\sigma(\{(1,0)\}|\theta) \geq \epsilon$ for some $\epsilon > 0$.

The structure of the proof is as follows. Step 1 begins by proving three preliminary properties of σ. Step 2 uses these properties to construct a candidate information structure and proves that it has a b-MIE. Step 3 proves that the introspective outcome implemented by the information structure approximates σ.

Step 1:

Lemma 6. *Take any $\tilde{\epsilon} \in (0, \epsilon)$ and define the introspective outcome $\tilde{\sigma}$ as follows:*

1. *For all $\theta \in \overline{\Theta}$, $\tilde{\sigma}(\cdot|\theta) = \frac{\sigma(\cdot|\theta) - \tilde{\epsilon}\delta_{(1,0)}}{1-\tilde{\epsilon}}$*

2. *For all $\theta \notin \overline{\Theta}$, $\tilde{\sigma}(\cdot|\theta) = \sigma(\cdot|\theta)$*

Then, for $\tilde{\epsilon}$ sufficiently small, $\tilde{\sigma}$ is (i) an introspective outcome, and (ii) satisfies b-obedience.

Proof. (i) holds because σ has the property that for all $\theta \in \overline{\Theta}$, $\sigma(\{(1,0)\}|\theta) \geq \epsilon$. (ii) holds because σ is assumed to satisfy b-obedience strictly. □

Lemma 7. *Suppose $\int_\Theta I_N d\tilde{\sigma}(I_A, I_N|\theta)dF(\theta) > 0$. For each $n > 1$, define*

$$\tilde{u}_n((I_A, I_N), \theta) \equiv \frac{1}{2n}\sum_{i=1}^{2n} u(I_A + \frac{(i-1)}{2n} I_N, \theta)$$

Then, there exists sufficiently large n such that

$$\int_\Theta \int_{\mathcal{I}} \tilde{u}_n((I_A, I_N), \theta) d\sigma(I_A, I_N|\theta)dF(\theta) > 0 \tag{25}$$

Proof. For each $n \in \mathbb{N}$, $\tilde{u}_n$ is the finite sum of integrable functions and is therefore integrable. Furthermore, since $u(I, \theta)$ is Riemann integrable, $\lim_{n\to\infty} \tilde{u}_n((I_A, I_N), \theta) = \int_0^{I_N} u(I_A + i, \theta)di$ for all (I_A, I_N, θ). Finally, each $\tilde{u}_n$ is dominated by the integrable function $\overline{u}$ defined pointwise by $\overline{u}((I_A, I_N), \theta) \equiv \max\{|\max_{\theta\in\Theta} u(1, \theta)|, |\inf_{\theta\in\Theta} u(0, \theta)|\}$. Thus, applying the Dominated Convergence Theorem,

$$\lim_{n\to\infty} \int_\Theta \int_{\mathcal{I}} \tilde{u}_n((I_A, I_N), \theta) d\sigma(I_A, I_N|\theta)dF(\theta) = \int_\Theta \int_{\mathcal{I}} \int_0^{I_N} u(I_A + i, \theta) di d\sigma(I_A, I_N|\theta)dF(\theta)$$

Finally, as (8) is strictly positive by assumption, (25) holds for large enough n. □

Now let $N = n$, where n is obtained from Lemma 7 above.

Lemma 8. *Given $\tilde{\epsilon} > 0$, there exists $\eta(\tilde{\epsilon}) \in (0, 1)$ such that (i) $\lim_{\tilde{\epsilon}\to 0} \eta(\tilde{\epsilon}) = 0$, (ii)*

$$\frac{\tilde{\epsilon}}{N-1} \int_{\overline{\Theta}} u(0, \theta)dF(\theta) + \eta(\tilde{\epsilon}) \int_{\Theta\setminus\overline{\Theta}} u(0, \theta)dF(\theta) \geq 0 \tag{26}$$

and, (iii) if $\int_\Theta I_N d\tilde{\sigma}(I_A, I_N|\theta)dF(\theta) > 0$, then

$$\int_\Theta \int_{\mathcal{I}} \frac{1}{N} \sum_{i=1}^{N} (1 - \eta(\tilde{\epsilon}))^{-i+1} u(I_A + I_N \frac{(i-1)}{N}, \theta) d\sigma(I_A, I_N|\theta)dF(\theta) \geq 0 \tag{27}$$

and

$$\int_{\Theta}\int_{\mathcal{I}}\frac{1}{N}\sum_{i=1}^{N}(1-\eta(\tilde{\epsilon}))^{-i+1}u(b(\theta),\theta)d\sigma(I_A,I_N|\theta)dF(\theta)<0 \tag{28}$$

Proof. $\underline{\eta}(\tilde{\epsilon}) \equiv \min\{\frac{1}{2}, \frac{\frac{\tilde{\epsilon}}{N-1}\int_{\overline{\Theta}}u(0,\theta)dF(\theta)}{-\int_{\Theta\setminus\overline{\Theta}}u(0,\theta)dF(\theta)}\} > 0$ is decreasing in $\tilde{\epsilon}$ and converges to zero. Thus, if $\int_{\Theta} I_N d\tilde{\sigma}(I_A, I_N|\theta)dF(\theta) = 0$, then setting $\eta(\tilde{\epsilon}) = \underline{\eta}(\tilde{\epsilon})$ yields the claim. Meanwhile, if $\int_{\Theta} I_N d\tilde{\sigma}(I_A, I_N|\theta)dF(\theta) > 0$, then since (25) holds, there exists $\overline{\eta} > 0$ such that $\eta < \overline{\eta}$ implies (27) holds under η, while since (9) holds strictly under σ, there exists $\tilde{\eta} > 0$ such that $\eta < \tilde{\eta}$ implies (28) holds under η. Then, letting $\eta(\tilde{\epsilon}) \equiv \min\{\underline{\eta}(\tilde{\epsilon}), \overline{\eta}, \tilde{\eta}\}$ yields the claim. □

Step 2: Take $\tilde{\epsilon} \in (0, \epsilon)$ sufficiently small so Lemma 6 is satisfied. For each $\theta \in \Theta$, consider the joint measure $\tilde{\pi}^{\tilde{\epsilon}}(\cdot|\theta) \in \Delta(\mathcal{I} \times \mathbb{Z}_+ \times \Delta(\mathbb{Z}_+ \cup \{\infty\}))$ defined as follows:

1. **Step 1:** Draw (I_A, I_N) according to $\sigma(\cdot|\theta)$
2. **Step 2:** Independently, draw a $z \in \mathbb{Z}_+$ with probability $\eta(\tilde{\epsilon})(1-\eta(\tilde{\epsilon}))^z$.
3. **Step 3a:** If $\theta \in \overline{\Theta}$ and $(I_A, I_N) = (1, 0)$ then draw $\mu_1 \in \Delta(\mathbb{Z}_+ \cup \{\infty\})$ with probability one, where

$$\mu_1(s) \equiv \begin{cases} \frac{\tilde{\epsilon}}{N-1}, & s \in \{1, ..., N-1\} \\ 1-\tilde{\epsilon}, & s = 0 \\ 0, & \text{otherwise} \end{cases}$$

4. **Step 3b:** If $\theta \notin \overline{\Theta}$ or $(I_A, I_N) \neq (1, 0)$, then draw $\mu_{(z,I_A,I_N)} \in \Delta(\mathbb{Z}_+ \cup \{\infty\})$ with probability one, where

$$\mu_{(z,I_A,I_N)}(s) \equiv \begin{cases} I_A, & s = 0 \\ \frac{I_N}{N}, & s \in \{z, ..., z+N-1\} \\ 1-I_A-I_N, & s = \infty \\ 0, & \text{otherwise} \end{cases}$$

Let $\pi^{\tilde{\epsilon}}(\cdot|\theta)$ denote the marginal distribution over $\mathbb{Z}_+ \cup \{\infty\}$ induced by $\tilde{\pi}^{\tilde{\epsilon}}(\cdot|\theta)$. Further let $\pi^{\tilde{\epsilon}}$ denote the joint measure induced by $(\pi^{\tilde{\epsilon}}(\cdot|\theta))_{\theta\in\Theta}$ and F over $(\mathbb{Z}_+ \cup \{\infty\}) \times$

$\Delta(\mathbb{Z}_+ \cup \{\infty\}) \times \Theta$, i.e., so $\pi^{\tilde{\epsilon}}(X \times Y \times Z) \equiv \int_Z \int_Y \int_X d\mu(s) d\pi^{\tilde{\epsilon}}(\cdot|\theta) dF(\theta)$ for all $X \in \mathbb{B}(\mathbb{Z}_+ \cup \{\infty\})$, $Y \in \mathbb{B}(\Delta(\mathbb{Z}_+ \cup \{\infty\}))$ and $Z \in \mathbb{B}(\Theta)$, $\pi_s^{\tilde{\epsilon}}(\cdot)$ denote the marginal distribution over $\mathbb{Z}_+ \cup \{\infty\}$, and $S \equiv \operatorname{supp}(\pi_s^{\tilde{\epsilon}}(\cdot))$. Finally, let $\mathcal{S}^{\tilde{\epsilon}} \equiv (S, (\pi^{\tilde{\epsilon}}(\cdot|\theta))_{\theta \in \Theta})$.

I will prove that (i) for all $s \leq N-1$, $\alpha^{\mathcal{S},b,1}(s) = \alpha^{\mathcal{S},b,2}(s) = 1$, (ii) for all $N \leq s < \infty$, $\alpha^{\mathcal{S},b,1}(s) = 0$ but $\alpha^{\mathcal{S},b,k}(s) = 1$ for $k = s$, and (iii) for $s = \infty$, $\alpha^{\mathcal{S},b,1}(s) = \alpha^{\mathcal{S},b}(s) = 0$. If so, then applying Lemma 5, every agent who invests under a signal $s = S$ at $k = 1$ also invests at $k = 2$, so $\mathcal{S}$ has a b-MIE. Furthermore, in this b-MIE (i) all agents observing $s \leq N-1$ are anchors, (ii) all agents observing $N \leq s < \infty$ are non-anchors, and (iii) all agents observing $s = \infty$ do not invest.

I split the proof into four parts.

Part 1: Signal $s = 0$

Proof. First, take any (I_A, I_N) drawn on state θ. An agent observes $s = 0$ with probability $(1-\tilde{\epsilon})I_A$ if $\theta \in \overline{\Theta}$ and $(I_A, I_N) = (1, 0)$, and probability I_A if $\theta \notin \overline{\Theta}$. Thus, the expected payoff of the agent from investing against the benchmark is proporitional to the left-hand side of (6) under $\tilde{\sigma}$ defined in Lemma 6. By Lemma 6, this is positive, so $\alpha^{\mathcal{S},b,1}(s) = 1$. Next, provided that (at least) a measure of I_A other agents are investing against the benchmark whenever an agent observes a signal of 0, the unconditional payoff of the agent from investing is at least (7) under $\tilde{\sigma}$. By a similar argument to the above, this is positive, so $\alpha^{\mathcal{S},b,2}(s) = 1$. □

Part 2: Signal $s = \infty$

Proof. Throughout, suppose $\int_\Theta (1 - I_A - I_N) d\tilde{\sigma}(I_A, I_N|\theta) dF(\theta) > 0$, for otherwise no such signal will be drawn. First consider $k = 1$. Fix (I_A, I_N) drawn on state θ. The probability that the agent observes $s = \infty$ is exactly $(1-\tilde{\epsilon})(1 - I_A - I_N)$ if $\theta \in \overline{\Theta}$ and $(I_A, I_N) = (1, 0)$, and $(1 - I_A - I_N)$ if $\theta \notin \overline{\Theta}$. Therefore, the agent's unconditional payoff from investing against the benchmark is a scalar multiple of (10) under $\tilde{\sigma}$, which by Lemma 6 is strictly negative. Thus, $\alpha^{\mathcal{S},b,1}(s) = 0$. Now take any $k > 1$. When (I_A, I_N) is drawn, the upper bound on the measure of agents who invest at $k-1$, i.e., with $\alpha^{k-1}(s) = 1$, is $1 - I_A - I_N$. Consequently, the agent's unconditional payoff from investing is at most a scalar multiple of the LHS of (11) under $\tilde{\sigma}$, which by Lemma 6 is strictly negative. Thus, $\alpha^{\mathcal{S},b,k}(s) = 0$. □

Part 3: Signals $s \in \{1, ..., N-1\}$

Proof. Take any $k \geq 1$. Fix (I_A, I_N) drawn on state θ. If $\theta \in \overline{\Theta}$, the agent observes signal $s \in \{1, ..., N-1\}$ with probability of at least $\frac{\tilde{\epsilon}}{N-1}$. If $\theta \notin \overline{\Theta}$, the probability that signal s is observed is at most $\eta(\tilde{\epsilon})$. Aggregating over all (I_A, I_N, θ) implies that regardless of other agents' behaviours, the agent's payoff from investing is at least (26), which is positive. Hence, $\alpha^{\mathcal{S},b,1}(s) = \alpha^{\mathcal{S},b,2}(s) = 1$. □

Part 4: Signals $N \leq s < \infty$

Proof. Throughout, suppose $\int_\Theta I_N d\sigma(I_A, I_N|\theta)dF(\theta) > 0$, i.e., non-anchors are drawn with strictly positive probability, for otherwise no such signal will be drawn.

I first prove that for all such s, $\alpha^{\mathcal{S},b,k}(s) = 1$ for $k = s$. Suppose k is observed by an agent when $((I_A, I_N), \theta)$ and $z \in \mathbb{Z}_+$ is drawn. This occurs if and only if $z \in \{k-(N-1), k\}$. Furthermore, by the induction hypothesis, the mass of other agents who invest are those who observes a $k' < k$, which is $I_A + I_N \frac{k-z}{N}$. Hence, the agent's payoff from investing is $u(I_A + I_N \frac{k-z}{N}, \theta)$. Since z is drawn with probability $\eta(\tilde{\epsilon})(1-\eta(\tilde{\epsilon}))^z$, aggregating over all relevant z and $((I_A, I_N), \theta)$ yields an unconditional expected payoff for the agent from investing of

$$\int_\Theta \int_{\mathcal{I}} \eta(\tilde{\epsilon}) \sum_{z=k-(N-1)}^{k} u(I_A + I_N \frac{k-z}{N}, \theta)(1-\eta(\tilde{\epsilon}))^z d\sigma(I_A, I_N|\theta)dF(\theta)$$

$$= \eta(\tilde{\epsilon})(1-\eta(\tilde{\epsilon}))^k \int_\Theta \int_{\mathcal{I}} \frac{1}{N} \sum_{i=1}^{N} (1-\eta(\tilde{\epsilon}))^{-i+1} u(I_A + \frac{(i-1)}{N}, \theta) d\sigma(I_A, I_N|\theta)dF(\theta) \geq 0$$

which implies (2) holds for $k = s$. Thus, $\alpha^{\mathcal{S},b,k}(s) = 1$.

From here, a similar argument to the above implies an agent's expected payoff from investing against the benchmark under such an s is proportional to (28), which is strictly negative. Thus, $\alpha^{\mathcal{S},b,1}(s) = 0$. □

Step 3: I now construct a b-implementable sequence $(\sigma^n)_{n\geq 1}$ which converges to σ. This implies σ is approximately b-implementable.

Take $\tilde{\epsilon} > 0$ which satisfies Lemma 6. Then, $\mathcal{S}^{\tilde{\epsilon}}$ defined in Step 2 is an introspective outcome with a b-MIE. Denote the introspective outcome implemented by it by $\sigma^{\tilde{\epsilon}}$.

I start by bounding the difference $|\sigma^{\tilde{\epsilon}}(\cdot|\theta) - \sigma(\cdot|\theta)|$. By Step 2, an agent who observes s is an anchor if and only if $s \leq N-1$, and a non-anchor if and only if

$N \leq s < \infty$. Hence, for all $W \in \mathbb{B}(\mathcal{I})$,

$$\begin{aligned}
\sigma^{\tilde{\epsilon}}(W|\theta) &= \pi^{\tilde{\epsilon}}\left(\left\{\mu : \begin{array}{c} \mu(\{s \leq N-1\}) = I_A, \text{ and} \\ \mu(\{N \leq s < \infty\}) = I_N \end{array}\right\}\middle|\theta\right) \\
&= \left(\begin{array}{c} Pr(z < N)\int_W \tilde{\pi}^{\tilde{\epsilon}}\left(\left\{\mu : \begin{array}{c} \mu(\{s \leq N-1\}) = I_A, \text{ and} \\ \mu(\{N \leq s < \infty\}) = I_N \end{array}\right\}\middle| z < N, (I_A, I_N), \theta\right) d\sigma(I_A, I_N|\theta) \\ +Pr(z \geq N)\int_W \tilde{\pi}^{\tilde{\epsilon}}\left(\left\{\mu : \begin{array}{c} \mu(\{s \leq N-1\}) = I_A, \text{ and} \\ \mu(\{N \leq s < \infty\}) = I_N \end{array}\right\}\middle| z \geq N, (I_A, I_N), \theta\right) d\sigma(I_A, I_N|\theta) \end{array}\right) \\
&= \left(\begin{array}{c} Pr(z < N)\int_W \tilde{\pi}^{\tilde{\epsilon}}\left(\left\{\mu : \begin{array}{c} \mu(\{s \leq N-1\}) = I_A, \text{ and} \\ \mu(\{N \leq s < \infty\}) = I_N \end{array}\right\}\middle| z < N, (I_A, I_N), \theta\right) d\sigma(I_A, I_N|\theta) \\ +Pr(z \geq N)\sigma(W|\theta) \end{array}\right)
\end{aligned}$$

where the second to third inequality holds as conditional on drawing $z \geq N$ and a pair (I_A, I_N), I_A agents observe $s = 0$ and I_N agents observe $N \leq s < \infty$ with probability one. Therefore, as $Pr(z < N) = 1 - (1 - \eta(\tilde{\epsilon}))^{N+1}$, the above implies

$$|\sigma^{\tilde{\epsilon}}(W|\theta) - \sigma(W|\theta)| \leq 2(1 - (1 - \eta(\tilde{\epsilon}))^{N+1}), \quad \forall W \in \mathbb{B}(\mathcal{I})$$

Given the above, observe that

$$\begin{aligned}
\left\|\sigma^{\tilde{\epsilon}} - \sigma\right\| &= \sup_{\tilde{\Theta} \in \mathbb{B}(\Theta), W \in \mathbb{B}(\mathcal{I})} \left| \int_{\tilde{\Theta}} (\sigma^{\tilde{\epsilon}}(W|\theta) - \sigma(W|\theta)) dF(\theta) \right| \\
&\leq \sup_{\tilde{\Theta} \in \mathbb{B}(\Theta)} \sup_{\theta \in \tilde{\Theta}, W \in \mathbb{B}(\mathcal{I})} |\sigma^{\tilde{\epsilon}}(W|\theta) - \sigma(W|\theta)| F(\tilde{\Theta}) \leq 2(1 - (1 - \eta(\tilde{\epsilon}))^{N+1})
\end{aligned}$$

Finally, Step 1 implies that $\eta(\tilde{\epsilon}) \to 0$ when $\tilde{\epsilon} \to 0$. Thus, $(\sigma^{\tilde{\epsilon}/n})_{n \geq 1}$ is a sequence of b-implementable introspective outcomes that converges to σ. □

A2.2. Weaker version of Theorem 2

I now prove Theorem 2 holds in general. First, suppose $\int_\Theta \int_{\mathcal{I}} I_N u(b(\theta), \theta) d\sigma(I_A, I_N|\theta) dF(\theta) > 0$. Define $\overline{\sigma}$ as follows.

$$\overline{\sigma}(\cdot|\theta) = \begin{cases} \delta_{(1/2,1/2)}, & u(0,\theta) > 0 \\ \delta_{(1,0)}, & u(0,\theta) \leq 0 \leq u(b(\theta),\theta) \\ \delta_{(0,0)}, & u(b(\theta),\theta) < 0 \text{ and } \theta \notin \underline{\Theta} \end{cases}, \quad \forall \theta \in \Theta \tag{29}$$

For all $n \in \mathbb{N}$, let $\sigma^n \equiv \frac{\tilde{\epsilon}}{2n}\overline{\sigma} + (1 - \frac{\tilde{\epsilon}}{2n})\sigma$, where $\tilde{\epsilon} > 0$ is chosen sufficiently small so (9) holds strictly. By construction, $\|\sigma - \sigma^n\| \leq \frac{\tilde{\epsilon}}{2n}$. Furthermore, σ^n satisfies the conditions of Appendix A2.1 strictly, so for each $n \in \mathbb{N}$, there exists a $\tilde{\sigma}^n$ which is b-implementable and satisfies $\|\tilde{\sigma}^n - \sigma^n\| < \tilde{\epsilon}/(2n)$. Combining these yields $\|\sigma - \tilde{\sigma}^n\| \leq \|\sigma - \sigma^n\| + \|\sigma^n - \tilde{\sigma}^n\| < \tilde{\epsilon}/n$, which implies the sequence $(\tilde{\sigma}^n)_{n\geq 1}$ converges to σ.

Next, suppose $\int_\Theta \int_\mathcal{I} I_N u(b(\theta), \theta) d\sigma(I_A, I_N|\theta) dF(\theta) = 0$. Define $\overline{\sigma}_2$ by setting $\overline{\sigma}_2(\cdot|\theta) \equiv \delta_{(1,0)}$ if $u(b(\theta), \theta) \geq 0$, and $\overline{\sigma}_2(\cdot|\theta) \equiv \delta_{(0,0)}$ otherwise. Then, by a similar argument to the above with $\overline{\sigma}_2$ in place of $\overline{\sigma}$, one obtains a sequence of b-implementable introspective outcomes that converges to σ. □

Appendix B: Other Proofs

Proof of Lemma 1 As the proof of Part 2 is simple, I only prove Part 1 here.

Take any information structure $\mathcal{S}$ with a b-MIE $\alpha^{\mathcal{S},b}$. I will show that for all $s \in S$, $\alpha^{\mathcal{S},b}(s)$ is a best-response for an agent against all other agents playing $\alpha^{\mathcal{S},b}$. I focus on the case where $\alpha^{\mathcal{S},b}(s) = 1$, i.e., the agent invests in the b-MIE under s, noting that the case with $\alpha^{\mathcal{S},b}(s) = 0$ is proven similarly.

As $\alpha^{\mathcal{S},b}(s) = 1$, there exists a $\overline{k} \geq 2$ such that for all $k \geq \overline{k}$, $\alpha^{\mathcal{S},b,k}(s) = 1$. That is,

$$\int_{\Delta(S)\times\Theta} u(I(\alpha^{\mathcal{S},b,k-1}|\mu), \theta) d\pi(\mu, \theta|s) \geq 0 \tag{30}$$

Meanwhile, notice that $(u(I(\alpha^{\mathcal{S},b,k-1}|\mu), \theta))_{k\geq\overline{k}}$ is a sequence of measurable functions which converges monotonically point-wise to $u(I(\alpha^{\mathcal{S},b}|\mu), \theta)$, and is bounded above and below by the integrable functions $u(1, \theta)$ and $u(0, \theta)$ respectively. Thus, by the Dominated Convergence Theorem,

$$\int_{\Delta(S)\times\Theta} u(I(\alpha^{\mathcal{S},b}|\mu), \theta) d\pi(\mu, \theta|s) = \lim_{k\to\infty} \int_{\Delta(S)\times\Theta} u(I(\alpha^{\mathcal{S},b,k-1}|\mu), \theta) d\pi(\mu, \theta|s) \tag{31}$$

Combining (30) and (31) implies $\int_{\Delta(S)\times\Theta} u(I(\alpha^{\mathcal{S},b}|\mu), \theta) d\pi(\mu, \theta|s) \geq 0$, so $\alpha^{\mathcal{S},b}(s) = 1$ is a best-response to other agents playing $\alpha^{\mathcal{S},b}$. □

Proof of Proposition 1. Take any upper b-obedient introspective outcome σ. For each $\theta \in \Theta$, define the transport map $\widetilde{T}_\theta : \mathcal{I} \to \mathcal{I}$ as follows:

1. If $u(I_A + I_N, \theta) \geq 0$, then $\widetilde{T}_\theta(I_A, I_N) \equiv (I_A, 1 - I_A)$. That is, it raises the mass of non-anchors drawn to $1 - I_A$.
2. If $u(I_A, I_N, \theta) < 0 \leq u(b(\theta), \theta)$, then $\widetilde{T}_\theta(I_A, I_N) \equiv (1, 0)$. That is, it raises the mass of anchors drawn to one, and reduces the mass of non-anchors to zero.
3. Otherwise, $\widetilde{T}_\theta(I_A, I_N) \equiv (I_A, I_N)$.

Let the introspective outcome $\tilde{\sigma}$ be defined such that $\tilde{\sigma}(\cdot|\theta)$ is the push-forward of $\sigma(\cdot|\theta)$ through $\widetilde{T}_\theta$ for all $\theta \in \Theta$. I first show that $\tilde{\sigma}$ is a candidate for an introspective outcome that satisfies the requirements of Corollary 1. That is, it investment and anchor-dominates σ, and satisfies all b-obedience constraints except possibly (9).

First, the maps $\{\widetilde{T}_\theta\}_{\theta \in \Theta}$ weakly raise the aggregate investment and mass of anchors on each pair drawn under σ, so $\tilde{\sigma}$ investment and anchor dominates σ. Next,

$$\begin{aligned}\int_\Theta \int_{\mathcal{I}} I_A u(b(\theta), \theta) d\tilde{\sigma}(I_A, I_N|\theta) dF(\theta) &= \left[\begin{array}{c} \int_\Theta \int_{\mathcal{I}} I_A u(b(\theta), \theta) d\sigma(I_A, I_N|\theta) dF(\theta) \\ + \int_\Theta \int_{(I_A, I_N): u(b(\theta), \theta) \geq 0} (1 - I_A) u(b(\theta), \theta) d\sigma(I_A, I_N|\theta) dF(\theta) \end{array}\right] \\ &\geq \int_\Theta \int_{\mathcal{I}} I_A u(b(\theta), \theta) d\sigma(I_A, I_N|\theta) dF(\theta) \geq 0\end{aligned}$$

and

$$\begin{aligned}\int_\Theta \int_{\mathcal{I}} I_A u(I_A, \theta) d\tilde{\sigma}(I_A, I_N|\theta) dF(\theta) &= \left[\begin{array}{c} \int_\Theta \int_{\mathcal{I}} I_A u(I_A, \theta) d\sigma(I_A, I_N|\theta) dF(\theta) \\ + \int_\Theta \int_{(I_A, I_N): u(b(\theta), \theta) \geq 0} (u(1, \theta) - I_A u(I_A, \theta)) d\sigma(I_A, I_N|\theta) dF(\theta) \end{array}\right] \\ &\geq \int_\Theta \int_{\mathcal{I}} I_A u(I_A, \theta) d\sigma(I_A, I_N|\theta) dF(\theta) \geq 0\end{aligned}$$

where the inequality holds as if $u(b(\theta), \theta) \geq 0$, then since $u(1, \theta) \geq u(b(\theta), \theta)$, $u(1, \theta) \geq I_A u(I_A, \theta)$ holds. Thus, $\tilde{\sigma}$ satisfies anchor-obedience. Then,

$$\begin{aligned}&\int_\Theta \int_{\mathcal{I}} \int_0^{I_N} u(I_A + i, \theta) di d\tilde{\sigma}(I_A, I_N|\theta) dF(\theta) \\ &= \left[\begin{array}{c} \int_\Theta \int_{\mathcal{I}} \int_0^{I_N} u(I_A + i, \theta) di d\sigma(I_A, I_N|\theta) dF(\theta) \\ + \int_\Theta \int_{(I_A, I_N): u(I_A + I_N, \theta) \geq 0} \int_{I_N}^{1 - I_A} u(I_A + i, \theta) di d\sigma(I_A, I_N|\theta) dF(\theta) \\ - \int_\Theta \int_{(I_A, I_N): u(I_A + I_N, \theta) < 0 \leq u(b(\theta), \theta)} \int_0^{I_N} u(I_A + i, \theta) di d\sigma(I_A, I_N|\theta) dF(\theta) \end{array}\right] \\ &\geq \int_\Theta \int_{\mathcal{I}} \int_0^{I_N} u(I_A + i, \theta) di d\sigma(I_A, I_N|\theta) dF(\theta) \geq 0\end{aligned}$$

so $\tilde{\sigma}$ satisfies non-anchor obedience. Finally,

$$\begin{aligned}&\int_\Theta\int_{\mathcal{I}}(1-I_A-I_N)u(b(\theta),\theta)d\tilde{\sigma}(I_A,I_N|\theta)dF(\theta)\\ =&\int_\Theta\int_{(I_A,I_N):u(I_A,I_N)<0\text{ and }u(b(\theta),\theta)<0}(1-I_A-I_N)u(b(\theta),\theta)d\sigma(I_A,I_N|\theta)dF(\theta)\leq 0\end{aligned}$$

while

$$\begin{aligned}&\int_\Theta\int_{\mathcal{I}}(1-I_A-I_N)u(I_A+I_N,\theta)d\tilde{\sigma}(I_A,I_N|\theta)dF(\theta)\\ =&\int_\Theta\int_{(I_A,I_N):u(I_A,I_N)<0\text{ and }u(b(\theta),\theta)<0}(1-I_A-I_N)u(I_A+I_N,\theta)d\sigma(I_A,I_N|\theta)dF(\theta)\leq 0\end{aligned}$$

so $\tilde{\sigma}$ satisfies (10) and (11) in downwards obedience.

Now, if (9) holds under $\tilde{\sigma}$, then $\tilde{\sigma}$ satisfies b-obedience and so satisfies the requirements of Corollary 1. If (9) does not hold, then let $\tilde{\sigma}_2$ be defined such that $\tilde{\sigma}_2(\cdot|\theta)$ is the pushforward of $\tilde{\sigma}(\cdot|\theta)$ through the map $\tilde{T}_2:\mathcal{I}\to\mathcal{I}$ defined by $\tilde{T}_2(I_A,I_N)=(I_A+I_N,0)$, i.e., $\tilde{\sigma}_2$ pools all non-anchors onto anchors. Clearly, $\tilde{\sigma}_2$ investment and anchor dominates σ, and it satisfies the first anchor, non-anchor and downwards obedience. As for the second anchor obedience constraint,

$$\begin{aligned}\int_\Theta\int_{\mathcal{I}}I_A u(I_A,\theta)d\tilde{\sigma}_2(I_A,I_N|\theta)dF(\theta)=&\int_\Theta\int_{\mathcal{I}}(I_A+I_N)u(I_A+I_N,\theta)d\tilde{\sigma}(I_A,I_N|\theta)dF(\theta)\\ \geq&\begin{pmatrix}\int_\Theta\int_{\mathcal{I}}I_A u(I_A,\theta)d\tilde{\sigma}(I_A,I_N|\theta)dF(\theta)\\ +\int_\Theta\int_{\mathcal{I}}\int_{I_A}^{I_A+I_N}u(i,\theta)did\tilde{\sigma}(I_A,I_N|\theta)dF(\theta)\end{pmatrix}\geq 0\end{aligned}$$

Thus, $\tilde{\sigma}_2$ is b-obedient, and so satisfies the requirements of Corollary 1 □

Proof of Lemma 2. Consider the following alternative problem (P')

$$\max_{\tilde{\sigma}\in\Delta(\mathcal{I}\times\Theta)}\int_{\mathcal{I}\times\Theta}v(I_A+I_N,\theta)d\tilde{\sigma}(I_A,I_N,\theta)\tag{32}$$

$$\text{s.t.}\quad\begin{pmatrix}\int_{\mathcal{I}\times\Theta}I_A u(b(\theta),\theta)d\tilde{\sigma}(I_A,I_N,\theta)\\ \int_{\mathcal{I}\times\Theta}I_A u(I_A,\theta)d\tilde{\sigma}(I_A,I_N,\theta)\\ \int_{\mathcal{I}\times\Theta}\int_{I_A}^{I_A+I_N}u(i,\theta)did\tilde{\sigma}(I_A,I_N,\theta)\end{pmatrix}\geq 0\tag{33}$$

$$\text{and}\quad\int_{\mathcal{I}\times\tilde{\Theta}}d\tilde{\sigma}(I_A,I_N,\theta)=\int_{\tilde{\Theta}}dF(\theta),\quad\forall\tilde{\Theta}\in\mathbb{B}(\Theta)\tag{34}$$

Any solution to (P), say σ, defines a joint measure that is feasible for (P') via $\tilde{\sigma}(W \times \tilde{\Theta}) \equiv \int_{\tilde{\Theta}} \int_W d\sigma(I_A, I_N|\theta) dF(\theta)$ for all $W \in \mathbb{B}(\mathcal{I})$ and $\tilde{\Theta} \in \mathbb{B}(\Theta)$. Likewise, because $\mathcal{I} \times \Theta$ is Polish, one can disintegrate any solution to (P') into the marginal F and an introspective outcome σ feasible for (P). Thus, any solution to (P') "solves" (P).

I now prove one solution to (P') exists, which proves Lemma 2. Throughout, I equip $\Delta(\mathcal{I} \times \Theta)$ with the weak* topology. I first prove the feasible set for (P') is compact. First, $\Delta(\mathcal{I} \times \Theta)$ is compact by Prokhorov's Theorem. Now take a sequence $(\tilde{\sigma}^n)_{n \geq 1}$ in $\Delta(\mathcal{I} \times \Theta)$ satisfying (33) and (34) that converges weakly to some $\tilde{\sigma} \in \Delta(\mathcal{I} \times \Theta)$. Since $u(I, \theta)$ is upper semicontinuous in (I, θ), the constraint function in (33) is upper semicontinuous, so $\tilde{\sigma}$ satisfies (33). Likewise, the constraint function in (34) is continuous, so $\tilde{\sigma}$ satisfies (34). Hence, the feasible set is compact. From here, $v(I, \theta)$ is upper semicontinuous, so the objective function in (32) is upper semicontinuous in $\tilde{\sigma}$. Hence, there exists a solution to (P'). □

Proof of Lemma 3. Take any feasible σ, so $\int_\Theta \int_{\mathcal{I}} U(I_A, I_N|\lambda, b, \theta) d\sigma(I_A, I_N|\theta) dF(\theta) \geq 0$ holds. Then,

$$\begin{aligned} V^D(\lambda^b, b) &= \int_\Theta \max_{(I_A, I_N)} \mathcal{L}(I_A, I_N|\lambda^b, b, \theta) dF(\theta) \\ &\geq \int_\Theta \int_{\mathcal{I}} \mathcal{L}(I_A, I_N|\lambda^b, b, \theta) d\sigma(I_A, I_N|\theta) dF(\theta) \\ &= \int_\Theta \int_{\mathcal{I}} (v(I_A + I_N, \theta) + \lambda^b U(I_A, I_N|b, \theta)) d\sigma(I_A, I_N|\theta) dF(\theta) \\ &\geq \int_\Theta \int_{\mathcal{I}} v(I_A + I_N, \theta) d\sigma(I_A, I_N|\theta) dF(\theta) = V(\sigma) \end{aligned}$$

Hence, $V^D(\lambda^b, b)$ is an upper bound on the designer's payoff in the relaxed problem. Notice then that the first inequality holds, if and only if (C1) holds, and the last inequality holds if and only if (C2) hold. This proves the claim. □

Proof of Lemma 4. First, take any $\theta \in \overline{\Theta}^b$ and $(I_A, I_N) \in \mathcal{I}$. Because $u(b(\theta), \theta)$ is non-negative and $I_A \leq 1 - I_N$, $I_A u(b(\theta), \theta) \leq (1 - I_N) u(b(\theta), \theta)$. Because $u(I, \theta)$ is non-decreasing in I, $u(i, \theta) \leq u(i + (1 - I_A - I_N), \theta)$. Therefore,

$$\mathcal{L}(I_A, I_N|\lambda^b, b, \theta) = v(I_A + I_N, \theta) + \lambda^b \left(I_A u(b(\theta), \theta) + \int_{I_A}^{I_A + I_N} u(i, \theta) di \right)$$

$$\leq v(1,\theta) + \lambda^b\bigg((1-I_N)u(b(\theta),\theta) + \int_{I_A}^{I_A+I_N} u(i+(1-I_A-I_N),\theta)di\bigg)$$
$$= v(1,\theta) + \lambda^b\bigg((1-I_N)u(b(\theta),\theta) + \int_{1-I_N}^{1} u(i,\theta)di\bigg) = \mathcal{L}(1-I_N, I_N|\lambda^b, b, \theta)$$

Lastly, a standard first-order approach shows $\mathcal{L}(1-I_N, I_N|\lambda^b, b, \theta)$ is maximized at $I_N = 1-b(\theta)$. Hence, $(b(\theta), 1-b(\theta)) \in \arg\max_{(I_A,I_N)\in\mathcal{I}} \mathcal{L}(1-I_N, I_N|\lambda^b, b, \theta)$. In turn, Condition 1 implies (C1) holds for all $\theta \in \overline{\Theta}^b$ under σ.

Next, take any $\theta \notin \overline{\Theta}^b$. Then, Assumption 1 implies $u(b(\theta),\theta) = u(0,\theta)$. Hence, for all (I_A, I_N),

$$\mathcal{L}(I_A, I_N|\lambda^b, b, \theta) = v(I_A+I_N,\theta) + \lambda^b\bigg(I_A u(0,\theta) + \int_{I_A}^{I_A+I_N} u(i,\theta)di\bigg)$$
$$\leq v(1,\theta) + \lambda^b \int_0^{I_A+I_N} u(i,\theta)di = \mathcal{L}(0, I_A+I_N|\lambda^b, b, \theta)$$

The above implies $\{(0,I) : I \in \arg\max_{I'\in[0,1]} \mathcal{L}(0, I'|\lambda^b, b, \theta)\} \subseteq \arg\max_{(I_A,I_N)\in\mathcal{I}} \mathcal{L}(1-I_N, I_N|\lambda^b, b, \theta)$. In turn, Condition 2 implies (C1) holds for all $\theta \notin \overline{\Theta}^b$ under σ. □

Proof of Proposition 4. Let $\mathcal{U}^b_{\underline{\theta}} \equiv \int_\Theta \int_{\mathcal{I}} U(I_A, I_N|\lambda, b, \theta)d\underline{\sigma}^b_{\underline{\theta}}(I_A, I_N|\theta)dF(\theta)$ denote agents' total expected perceived payoffs from investing under $\underline{\sigma}^b_{\underline{\theta}}$. By the continuity of F, $\mathcal{U}^b_{\underline{\theta}}$ is continuous in $\underline{\theta}$.

By the in-text discussion, to prove Proposition 4, one must show that $\underline{\sigma}^b$ is feasible and satisfies (C2). There are two cases to consider, depending on λ^b.

First, suppose $\lambda^b = 0$. Then, any $\underline{\sigma}^b_{\underline{\theta}}$ satisfies (C2). Meanwhile, by Theorem 3 of Milgrom and Segal (2002) (henceforth "Envelope Theorem"), the right derivative of $V^D(\cdot, b)$ with respect to b evaluated at $\lambda = 0$ is

$$(V^D)'_+(0,b) = \int_{\overline{\Theta}^b} U(b(\theta), 1-b(\theta)|\lambda, b, \theta)dF(\theta) + \int_{\Theta\setminus\overline{\Theta}^b} \sup\bigg\{\int_0^I u(i,\theta)di : v(1,\theta) = v(I,\theta)\bigg\}dF(\theta)$$
$$= \int_{\overline{\Theta}^b} U(b(\theta), 1-b(\theta)|\lambda, b, \theta)dF(\theta) + \int_{\theta\in\Theta\setminus\overline{\Theta}^b:\theta\geq\theta_0} \int_0^1 u(i,\theta)didF(\theta) = \mathcal{U}^b_{\theta_0}$$

Since the above must be non-negative for $\lambda = 0$ to minimize $V^D(\cdot, b)$, $\mathcal{U}^b_{\theta_0} \geq 0$. Hence, $\underline{\sigma}^b_{\theta_0}$ is feasible. In turn, $\underline{\sigma}^b$ is feasible.

Next, suppose $\lambda^b > 0$. Observe then that for all $\theta \in \Theta$, $v(I^b_-(\theta)\theta) \leq v(I^b_+(\theta),\theta)$. Therefore, $U(0, I^b_-(\theta)|b,\theta) \geq U(0, I^b_+(\theta)|b,\theta)$ must hold. In turn, $\mathcal{U}^b_{\underline{\theta}}$ is non-decreasing in $\underline{\theta}$. Furthermore, the Envelope Theorem implies the left derivative of $V^D(\lambda, b)$ at $\lambda = \lambda^b$ satisfies $(V^D)'_-(\lambda^b, b) = \mathcal{U}_0^b \leq 0$, while $(V^D)'_+(\lambda^b, b) = \mathcal{U}_1^b \geq 0$. Together, these imply $\mathcal{U}^b_{\underline{\theta}^b} = 0$, so $\underline{\sigma}^b$ is both feasible and satisfies (C2). □

Proof of Proposition 5. I first show that under σ^b, anchor obedience holds. The first constraint holds by the definition of $\overline{\theta}^b$. The second constraint holds as

$$\int_\Theta \int_{\mathcal{I}} I_A u(I_A, \theta) d\sigma^b(I_A, I_N|\theta) dF(\theta) = \begin{pmatrix} \int_{\overline{\Theta}^b} b(\theta) u(b(\theta), \theta) dF(\theta) \\ \int_{[\overline{\theta}^b, 1]\setminus\overline{\Theta}^b} \min\{\underline{I}(\theta), I^b(\theta)\} u(\min\{\underline{I}(\theta), I^b(\theta)\}, \theta) dF(\theta) \end{pmatrix}$$

$$\geq \begin{pmatrix} \int_{\overline{\Theta}^b} b(\theta) u(b(\theta), \theta) dF(\theta) \\ \int_{[\overline{\theta}^b, 1]\setminus\overline{\Theta}^b} \min\{\underline{I}(\theta), I^b(\theta)\} u(0, \theta) dF(\theta) \end{pmatrix} \geq 0$$

I now show that non-anchor obedience (8) holds. There are two cases to consider.

First, suppose anchors' perceived payoffs from investing under σ^b is strictly positive. Then $\overline{\theta}^b = 0$, so $\sigma^b = \sigma_0^b$. In turn, the left-hand side of (8) under σ^b is

$$\int_{\overline{\Theta}^b} \int_{b(\theta)}^1 u(i, \theta) di dF(\theta) + \int_{\Theta\setminus\overline{\Theta}^b} \int_{\min\{I^b(\theta), \underline{I}(\theta)\}}^{I^b(\theta)} u(i, \theta) dF(\theta)$$

Because $u(b(\theta), \theta) \geq 0$ on all $\theta \in \overline{\Theta}^b$, and $u(i, \theta) \geq 0$ for all $i \geq \underline{I}(\theta)$, the term above is positive. Thus, non-anchor obedience holds.

Next, suppose anchors' perceived payoffs from investing under σ^b is zero. Then, the left-hand side of (8) under σ^b is equal to the total perceived payoffs from investment under σ^b, which is $\mathcal{U}^b_{\underline{\theta}^b} \geq 0$. Thus, non-anchor obedience holds. □

Proof of Propositions 6 and 7. I prove these results through a series of claims.

Claim 1. *For all $\theta \in \overline{\Theta}^b$, (i) $I^b(\theta) \geq I^{\tilde{b}}(\theta)$ and (ii) $I_A^b(\theta) \geq I_A^{\tilde{b}}(\theta)$*

Proof. (i) holds as $I^b(\theta) = 1$. For (ii), there are two cases to consider. First, suppose $\theta \in \overline{\Theta}^{\tilde{b}}$. Then, $I_A^b(\theta) = b(\theta)$ and $I_A^{\tilde{b}}(\theta) = \tilde{b}(\theta)$. Since $b \geq \tilde{b}$, $I_A^b(\theta) \geq I_A^{\tilde{b}}(\theta)$. Next, suppose $\theta \in \overline{\Theta}^b\setminus\overline{\Theta}^{\tilde{b}}$. Then, $I_A^b(\theta) = b(\theta)$ and $I_A^{\tilde{b}}(\theta) \leq \underline{I}(\theta)$. Since $u(b(\theta), \theta) \geq 0$, $b(\theta) \geq \underline{I}(\theta)$ holds. Hence, $I_A^b(\theta) \geq I_A^{\tilde{b}}(\theta)$. □

Claim 2. *For any $\theta \in \overline{\Theta}^b\setminus\overline{\Theta}^{\tilde{b}}$, (i) $U(\tilde{b}(\theta), 1 - \tilde{b}(\theta)|\tilde{b}, \theta) \leq U(b(\theta), 1 - b(\theta)|b, \theta)$ and (ii) $\int_0^{I^{\tilde{b}}(\theta)} u(i, \theta) di \leq U(b(\theta), 1 - b(\theta)|b, \theta)$.*

Proof. Part (i) is easy to verify. As for Part (ii),

$$\begin{aligned}\int_0^{I^{\tilde{b}}(\theta)} u(i,\theta)di &= \min\{\underline{I}(\theta), I^{\tilde{b}}(\theta)\}u(0,\theta) + \int_{\min\{\underline{I}(\theta),I^{\tilde{b}}(\theta)\}}^{I^b(\theta)} u(i,\theta)di \\ &\leq \min\{\underline{I}(\theta), I^{\tilde{b}}(\theta)\}u(b(\theta),\theta) + \int_{\min\{\underline{I}(\theta),I^{\tilde{b}}(\theta)\}}^{I^b(\theta)} u(i,\theta)di \\ &\leq \min\{\underline{I}(\theta), I^{\tilde{b}}(\theta)\}u(b(\theta),\theta) + \int_{\min\{\underline{I}(\theta),I^{\tilde{b}}(\theta)\}}^{1} u(i,\theta)di \\ &\leq b(\theta)u(b(\theta),\theta) + \int_{b(\theta)}^{1} u(i,\theta)di = U(b(\theta), 1-b(\theta)|b,\theta)\end{aligned}$$

where the last two inequalities follow from similar arguments to Lemma 4's proof. □

Claim 3. *If $\lambda^{\tilde{b}} = 0$, then for all $\theta \notin \overline{\Theta}^b$, $I^b(\theta) \geq I^b(\tilde{\theta})$.*

Proof. I first show that $\underline{\theta}^b \leq \underline{\theta}^{\tilde{b}}$. To start, applying the Envelope Theorem,

$$\begin{aligned}(V^D)'_+(0,b) &= \int_{\overline{\Theta}^b} U(b(\theta), 1-b(\theta)|\lambda, b, \theta)dF(\theta) + \int_{\theta\in\Theta\backslash\overline{\Theta}^b:\theta\geq\theta_0}\int_0^1 u(i,\theta)didF(\theta) \\ &\geq \int_{\overline{\Theta}^{\tilde{b}}} U(\tilde{b}(\theta), 1-\tilde{b}(\theta)|\lambda, \tilde{b}, \theta)dF(\theta) + \int_{\theta\in\Theta\backslash\overline{\Theta}^{\tilde{b}}:\theta\geq\theta_0}\int_0^1 u(i,\theta)didF(\theta) \\ &= (V^D)'_+(0,\tilde{b}) \geq 0\end{aligned}$$

where the inequality holds because of Claim 2. Hence, $\lambda = 0$ also minimizes $V^D(\lambda, b)$, so $\lambda^b = 0$. This implies that for all $\theta \notin \overline{\Theta}^b$, $I^b_-(\theta) = I^{\tilde{b}}_-(\theta)$ and $I^b_+(\theta) = I^{\tilde{b}}_+(\theta)$. Therefore, under the threshold $\underline{\theta} = \underline{\theta}^{\tilde{b}}$,

$$\begin{aligned}\mathcal{U}^b_{\underline{\theta}^{\tilde{b}}} &= \begin{pmatrix} \int_{\overline{\Theta}^{\tilde{b}}} U(b(\theta), 1-b(\theta)|b,\theta)dF(\theta) + \int_{\overline{\Theta}^b\backslash\overline{\Theta^{\tilde{b}}}} U(b(\theta), 1-b(\theta)|b,\theta)dF(\theta) \\ + \int_{\theta\in\Theta\backslash\overline{\Theta}^b:\theta\geq\underline{\theta}^{\tilde{b}}}\int_0^{I^b_+(\theta)} u(i,\theta)didF(\theta) + \int_{\theta\in\Theta\backslash\overline{\Theta}^b:\theta<\underline{\theta}^{\tilde{b}}}\int_0^{I^b_-(\theta)} u(i,\theta)didF(\theta) \end{pmatrix} \\ &\geq \begin{pmatrix} \int_{\overline{\Theta}^{\tilde{b}}} U(\tilde{b}(\theta), 1-\tilde{b}(\theta)|\tilde{b},\theta)dF(\theta) + \int_{\overline{\Theta}^b\backslash\overline{\Theta^{\tilde{b}}}}\int_0^{I^{\tilde{b}}} u(i,\theta)didF(\theta) \\ + \int_{\theta\in\Theta\backslash\overline{\Theta}^b:\theta\geq\underline{\theta}^{\tilde{b}}}\int_0^{I^{\tilde{b}}_+(\theta)} u(i,\theta)didF(\theta) + \int_{\theta\in\Theta\backslash\overline{\Theta}^b:\theta<\underline{\theta}^{\tilde{b}}}\int_0^{I^{\tilde{b}}_-(\theta)} u(i,\theta)didF(\theta) \end{pmatrix} \\ &= \mathcal{U}^{\tilde{b}}_{\underline{\theta}^{\tilde{b}}} \geq 0\end{aligned}$$

where the first inequality follows from Claim 2. This means $\underline{\theta}^b \leq \underline{\theta}^{\tilde{b}}$.

Now take any $\theta \in \Theta\backslash\overline{\Theta}^b$. If $I^{\tilde{b}}(\theta) = I^{\tilde{b}}_-(\theta)$, then $I^b(\theta) \geq I^b_-(\theta) = I^{\tilde{b}}_-(\theta) = I^{\tilde{b}}(\theta)$. If $I^{\tilde{b}}(\theta) = I^{\tilde{b}}_+(\theta)$, then $\theta \geq \underline{\theta}^{\tilde{b}}$. Since $\underline{\theta}^b \leq \underline{\theta}^{\tilde{b}}$, this means $\theta \geq \underline{\theta}^b$. Therefore, $I^b(\theta) =$

$I^b_+(\theta) = I^{\tilde{b}}_+(\theta) = I^{\tilde{b}}(\theta)$. □

Claim 4. *If $\lambda^{\tilde{b}} > 0$, then for all $\theta \notin \overline{\Theta}^b$, $I^b(\theta) \geq I^b(\tilde{\theta})$.*

Proof. First, notice

$$\begin{aligned}
(V^D)'_-(\lambda^{\tilde{b}}, b) &= \left(\begin{array}{c} \int_{\overline{\Theta}^{\tilde{b}}} U(b(\theta), 1-b(\theta)|b,\theta)dF(\theta) + \int_{\overline{\Theta}^b \backslash \overline{\tilde{\Theta}^b}} U(b(\theta), 1-b(\theta)|b,\theta)dF(\theta) \\ + \int_{\theta\in\Theta\backslash\overline{\Theta}^b} \int_0^{I^{\tilde{b}}_-(\theta)} u(i,\theta)didF(\theta) \end{array} \right) \\
&\geq \left(\begin{array}{c} \int_{\overline{\Theta}^{\tilde{b}}} U(\tilde{b}(\theta), 1-\tilde{b}(\theta)|\tilde{b},\theta)dF(\theta) + \int_{\overline{\Theta}^b \backslash \overline{\tilde{\Theta}^b}} \int_0^{I^{\tilde{b}}} u(i,\theta)didF(\theta) \\ + \int_{\theta\in\Theta\backslash\overline{\Theta}^b} \int_0^{I^{\tilde{b}}_-(\theta)} u(i,\theta)didF(\theta) \end{array} \right) \\
&= (V^D)'_-(\lambda^{\tilde{b}}, \tilde{b})
\end{aligned}$$

where the inequality follows from Claim 2. Applying a similar argument establishes $(V^D)'_+(\lambda^{\tilde{b}}, b) \geq (V^D)'_+(\lambda^{\tilde{b}}, \tilde{b})$. Hence, $\lambda^b \leq \lambda^{\tilde{b}}$.

Now if $\lambda^b = \lambda^{\tilde{b}}$, then an identical argument to Claim 3 proves Claim 4 holds. Meanwhile, if $\lambda^b < \lambda^{\tilde{b}}$, then for all $\theta \notin \overline{\Theta}^b$, $I^b_-(\theta) \geq I^{\tilde{b}}_+(\theta)$. That $I^b(\theta) \geq I^b_-(\theta)$ and $I^{\tilde{b}}_+(\theta) \geq I^{\tilde{b}}(\theta)$ holds on all such states then implies Claim 4 holds. □

Proof of Proposition 8 I will prove that $I^b(\theta) \in \{0,1\}$ on all non b-agreement states $\theta \notin \overline{\Theta}^b$. Since $I^b(\theta) = 1$ on all b-agreement states, this then implies σ^b perfectly coordinates investments on all states.

Take any non b-agreement state θ. Restricted convexity implies that for all $\lambda \geq 0$, either $\mathcal{L}(0, I|\lambda, b) \leq \mathcal{L}(0,1|\lambda,b)$ if $\int_0^I u(i,\theta)di \leq \int_0^1 u(i,\theta)di$, or $\mathcal{L}(0,I|\lambda,b) \leq \frac{\int_0^I u(i,\theta)di}{\int_0^1 u(i,\theta)di}\mathcal{L}(0,1|\lambda,b) + (1 - \frac{\int_0^I u(i,\theta)di}{\int_0^1 u(i,\theta)di})\mathcal{L}(0,0|\lambda,b)$ if $\int_0^I u(i,\theta)di > \int_0^1 u(i,\theta)di$. These imply either (i) $\mathcal{L}(0,0|\lambda^b,\theta) = \mathcal{L}(0,1|\lambda^b,\theta)$, so $I^b_-(\theta) = 0$ and $I^b_+(\theta) = 1$, (ii) $\mathcal{L}(0,0|\lambda^b,\theta) > \mathcal{L}(0,1|\lambda^b,\theta)$, so $I^b_-(\theta) = I^b_+(\theta) = 0$ or (iii) $\mathcal{L}(0,0|\lambda^b,\theta) < \mathcal{L}(0,1|\lambda^b,\theta)$, so $I^b_-(\theta) = I^b_+(\theta) = 1$. Since $I^b(\theta) \in \{I^b_-(\theta), I^b_+(\theta)\}$ by definition, $I_b(\theta) \in \{0,1\}$ holds. □

Proof of Corollary 2. Follows from in-text discussion. □

Proof of Corollary 3. Proposition 7 has already shown that $I^b_A(\theta) \geq I^{\tilde{b}}_A(\theta)$ for all $\theta \in \overline{\Theta}^b$. Meanwhile, by definition, $\overline{\theta}^b$ is non-increasing in b. Hence, if $\theta \in [\overline{\theta}^b, 1]\backslash\overline{\Theta}^b$, then $I^b_A(\theta) = \underline{I}(\theta) \geq I^{\tilde{b}}(\theta)$. Meanwhile, if $\theta \in [0, \overline{\theta}^b)\backslash\overline{\Theta}^b$, then $I^b_A(\theta) = I^{\tilde{b}}_A(\theta) = 0$. □

Proof of Corollary 4. Follows from in-text discussion. □

Proof of Proposition 9. As Part 2 follows from the in-text discussion, I only prove Part 1. The change in agents' expected perceived payoffs under the optimal introspective outcome from raising the benchmark from $\tilde{b}$ to b is

$$\int_{[\underline{\theta}^{\tilde{b}},1]\cup\overline{\Theta}^{b}}[U(b(\theta),1-b(\theta)|b,\theta)-U(\tilde{b}(\theta),1-\tilde{b}(\theta)|\tilde{b},\theta)]dF(\theta)$$
$$+\int_{([\underline{\theta}^{b},\underline{\theta}^{\tilde{b}}]\cup\overline{\Theta}^{b})\setminus\overline{\Theta}^{\tilde{b}}}U(I_A^b(\theta),1-I_A^b(\theta)|\tilde{b},\theta)dF(\theta)$$

Using the definition of $\underline{\theta}^{b}$ in (20), the sum is weakly positive. Meanwhile, for all $\theta\in[\underline{\theta}^{\tilde{b}},1]\cup\overline{\Theta}^{b}$, $u(b(\theta),\theta)=u(\tilde{b}(\theta),\theta)$, so $U(b(\theta),1-b(\theta)|b,\theta)=U(\tilde{b}(\theta),1-\tilde{b}(\theta)|\tilde{b},\theta)$. Hence, the first term is zero. Following the in-text argument then proves Part 1. □

Appendix C: Public Information Design

Optimal Public Information Structure. I begin by characterizing the optimal public information structure. As a first step, the next result characterize the unique b-MIE under any public information structure.

Lemma 9. *For all $b\in\mathcal{B}$, every public information structure has a unique b-MIE. In it, an agent invests under a signal if and only if he invests against the benchmark.*

Proof. Take a public information structure $\mathcal{S}$ and $b\in\mathcal{B}$. To prove the claim, it suffices to show that for all $s\in\mathcal{S}$, (i) $\alpha^{\mathcal{S},b,1}(s)=1\Rightarrow\alpha^{\mathcal{S},b,k}(s)=1$ for all $k\geq 1$, and (ii) $\alpha^{\mathcal{S},b,1}(s)=0\Rightarrow\alpha^{\mathcal{S},b,k}(s)=0$ for all $k\geq 1$. For (i), note that if $\alpha^{\mathcal{S},b,1}(s)=1$, then

$$\int_{\Delta(S)\times\Theta}u(I(\mu|\alpha^{\mathcal{S},b,1}),\theta)d\pi(\mu,\theta|s)=\int_{\Theta}u(1,\theta)d\pi(\theta|s)\geq\int_{\Theta}u(b(\theta),\theta)d\pi(\theta|s)\geq 0$$

so $\alpha^{\mathcal{S},b,2}(s)=1$. Repeating the argument inductively implies $\alpha^{k}(s)=1$ for all $k\geq 2$. For (ii), if $\alpha^{1}(s)=0$, then

$$\int_{\Delta(S)\times\Theta}u(I(\mu|\alpha^{\mathcal{S},b,1}),\theta)d\pi(\mu,\theta|s)=\int_{\Theta}u(0,\theta)d\pi(\theta|s)\leq\int_{\Theta}u(b(\theta),\theta)d\pi(\theta|s)<0 \quad (35)$$

so $\alpha^{\mathcal{S},b,2}(s)=0$. Repeating the argument implies $\alpha^{k}(s)=0$ for all $k\geq 2$. □

Lemma 9 implies under any signal observed, agents' behaviours are identical, and whether any agent invests depends on whether investing is optimal against the benchmark. That is, the public information design problem reduces to the problem of persuading all agents to invest, where these agents invest if and only if the payoff from investing against the benchmark is non-negative. Denoting the probability all agents invest on each state by $\tilde{\sigma} : \Theta \to [0,1]$, this problem is:

$$\max_{\tilde{\sigma}} \int_{\Theta} v(1,\theta)\tilde{\sigma}(\theta)dF(\theta) \quad \text{s.t.} \quad \int_{\Theta} u(b(\theta),\theta)\tilde{\sigma}(\theta)dF(\theta) \geq 0 \tag{Pub}$$

Intuitively, the designer finds it optimal to induces investment on all agreement states $[\overline{\theta}, 1]$. Meanwhile, because non-agreement states $\theta < \overline{\theta}$ are ranked by the "cost-to-benefit" ratio $R(\theta) \equiv -u(b(\theta),\theta)/v(1,\theta) = -\underline{u}(\theta)/v(1,\theta)$, the designer optimally induces investment on the highest non agreement states. Formally,

Theorem 3. *Let*

$$\tilde{\theta}^b \equiv \inf\left\{\theta \in [0,\overline{\theta}] : \int_{[\overline{\theta},1]} u(b(\hat{\theta}),\hat{\theta})dF(\hat{\theta}) + \int_{[\theta,\overline{\theta}]} \underline{u}(\hat{\theta})dF(\hat{\theta}) \geq 0\right\}$$

Then, $\tilde{\sigma}^b$ defined below solves the public information design problem:

$$\tilde{\sigma}^b(\theta) \equiv \begin{cases} 1, & \theta \geq \tilde{\theta}^b \\ 0, & \theta < \tilde{\theta}^b \end{cases}$$

Proof. Without loss, I assume that $\tilde{\theta}^b > 0$, for otherwise the claim immediately holds. This implies $R^b \equiv R(\tilde{\theta}^b) > 0$ and $\int_{[\overline{\theta},1]} u(b(\hat{\theta}),\hat{\theta})dF(\hat{\theta}) + \int_{[\tilde{\theta}^b,\overline{\theta}]} \underline{u}(\hat{\theta})dF(\hat{\theta}) = 0$. Now observe that (Pub), which is a linear programming problem, admits the following *dual problem*: choose a $\lambda \geq 0$ and a measurable function $\phi : \Theta \to \mathbb{R}_+$ to solve

$$\min_{\lambda,\phi} \int_{\Theta} \phi(\theta)dF(\theta) \quad \text{s.t.} \quad v(1,\theta) + \lambda u(b(\theta),\theta) \leq \phi(\theta), \quad \forall \theta \in \Theta$$

Given this, let $\lambda^* \equiv 1/R^b$ and $\phi^*(\theta) \equiv \max\{0, v(1,\theta) + \lambda^* u(b(\theta),\theta)\}$. Clearly, (λ^*, ϕ^*) is feasible for the dual problem. Furthermore, observe that $\phi^*(\theta) = v(1,\theta) + \lambda^* u(b(\theta),\theta)$ if $\theta \geq \tilde{\theta}^b$, and $\phi^*(\theta) = 0$ otherwise. Hence,

$$\underbrace{\int_{\Theta} \phi^*(\theta)dF(\theta)}_{\text{Value of dual objective under } (\lambda^*,\phi^*)} = \underbrace{\int_{[\tilde{\theta}^b,1]} v(1,\theta)dF(\theta)}_{\text{Value of primal objective under } \tilde{\sigma}^b}$$

$$+\lambda^* \underbrace{\int_{[\overline{\theta},1]} u(b(\hat{\theta}),\hat{\theta})dF(\hat{\theta}) + \int_{[\tilde{\theta}^b,\overline{\theta}]} \underline{u}(\hat{\theta})dF(\hat{\theta})}_{=0}$$

A standard Weak Duality argument then implies $\tilde{\sigma}^b$ solves the designer's problem. □

Proofs for Section 6. I start with three observations. First, under Assumptions 2, 4, 5 and 6, the optimal introspective outcome σ^b, which is characterized by Corollary 2, has all agents invest only on states $\theta \in [\underline{\theta}^b, 1]$, where

$$\underline{\theta}^b \equiv \min\left\{\theta \in [0,\overline{\theta}] : \int_{[\overline{\theta},1]} U(b(\hat{\theta}), 1-b(\hat{\theta})|b,\hat{\theta})dF(\hat{\theta}) + \int_{[\theta,\overline{\theta}]} \underline{u}(\hat{\theta})dF(\hat{\theta}) \geq 0\right\}$$

Second, because the designer is always weakly worse off under public information, $\tilde{\theta}^b \geq \underline{\theta}^b$ holds. Finally, both $\tilde{\theta}^b$ and $\underline{\theta}^b$ are non-increasing in the benchmark b.

Proof of Proposition 10. It suffices to show that under the stated conditions, $\tilde{\theta}^b > \underline{\theta}^b$. First, since $u(b(\theta),\theta) < u(1,\theta)$ for all $\theta \in \tilde{\Theta}$ and $F(\tilde{\Theta}) > 0$,

$$\int_{[\overline{\theta},1]} u(b(\theta),\theta)dF(\theta) + \int_{[}[0,\overline{\theta}]]\underline{u}(\theta)dF(\theta) < \int_{\Theta} u(1,\theta)dF(\theta) \leq 0$$

Hence, $\tilde{\theta}^b > 0$ and $\int_{[\overline{\theta},1]} u(b(\theta),\theta)dF(\theta) + \int_{[\tilde{\theta}^b,\overline{\theta}]} \underline{u}(\theta)dF(\theta) = 0$. Meanwhile,

$$u(b(\theta),\theta) < b(\theta)u(b(\theta),\theta) + \int_{b(\theta)}^{1} u(i,\theta)di = U(b(\theta), 1-b(\theta)|b,\theta), \quad \forall\theta \in \tilde{\Theta}$$

And so

$$\begin{array}{c} \int_{[\overline{\theta},1]} U(b(\theta), 1-b(\theta)|b,\theta)dF(\theta) \\ + \int_{[\tilde{\theta}^b,\overline{\theta}]} \underline{u}(\theta)dF(\theta) \end{array} > \begin{array}{c} \int_{[\overline{\theta},1]} U(b(\theta), 1-b(\theta)|b,\theta)dF(\theta) \\ + \int_{[\tilde{\theta}^b,\overline{\theta}]} \underline{u}(\theta)dF(\theta) \end{array} = 0$$

Therefore, $\underline{\theta}^b < \tilde{\theta}^b$. □

Proof of Proposition 11. I focus on when (i) $\tilde{\theta}^b > 0$ and (ii) $\underline{\theta}^b < \underline{\theta}^{\tilde{b}}$, otherwise Proposition 11 immediately holds. (i) implies $\int_{[\overline{\theta},1]} u(\tilde{b}(\hat{\theta}),\hat{\theta})dF(\hat{\theta}) + \int_{[\tilde{\theta}^b,\overline{\theta}]} \underline{u}(\hat{\theta})dF(\hat{\theta}) =$

0 and $\int_{[\overline{\theta},1]} u(b(\hat{\theta}),\hat{\theta})dF(\hat{\theta}) + \int_{[\tilde{\theta}^b,\overline{\theta}]} \underline{u}(\hat{\theta})dF(\hat{\theta}) = 0$. Combining these yield

$$\int_{\tilde{\theta}^{\tilde{b}}}^{1} [u(b(\theta),\theta) - u(\tilde{b}(\theta),\theta)]dF(\theta) + \int_{\tilde{\theta}^b}^{\tilde{\theta}^{\tilde{b}}} \underline{u}(\theta)dF(\theta) = 0 \tag{36}$$

Meanwhile (ii) implies that $\int_{[\overline{\theta},1]} U(\tilde{b}(\theta), 1-\tilde{b}(\theta)|\tilde{b},\theta)dF(\theta) + \int_{[\underline{\theta}^{\tilde{b}},\overline{\theta}]} \underline{u}(\theta)dF(\theta) = 0$ and $\int_{[\overline{\theta},1]} U(b(\theta),1-b(\theta)|b,\theta)dF(\theta) + \int_{[\underline{\theta}^b,\overline{\theta}]} \underline{u}(\theta)dF(\theta) \geq 0$, which combined yields

$$\int_{[\overline{\theta},1]} [U(b(\theta),1-b(\theta)|b,\theta) - U(\tilde{b}(\theta),1-\tilde{b}(\theta)|b,\theta)]dF(\theta) + \int_{\underline{\theta}^b}^{\underline{\theta}^{\tilde{b}}} \underline{u}(\theta)dF(\theta) \geq 0 \tag{37}$$

Further notice that

$$\begin{aligned}
&\int_{[\overline{\theta},1]} [U(b(\theta),1-b(\theta)|b,\theta) - U(\tilde{b}(\theta),1-\tilde{b}(\theta)|b,\theta)]dF(\theta) \\
&\quad = \int_{[\overline{\theta},1]} \left(b(\theta)u(b(\theta),\theta) - \tilde{b}(\theta)u(\tilde{b}(\theta),\theta) - \int_{\tilde{b}(\theta)}^{b(\theta)} u(i,\theta)di \right) dF(\theta) \\
&\quad \leq \int_{[\overline{\theta},1]} [u(b(\theta),\theta) - u(\tilde{b}(\theta),\theta)]dF(\theta)
\end{aligned}$$

Combining the above with (36) and (37) yields $-\int_{\underline{\theta}^b}^{\underline{\theta}^{\tilde{b}}} \underline{u}(\theta)dF(\theta) + \int_{\tilde{\theta}^b}^{\tilde{\theta}^{\tilde{b}}} \underline{u}(\theta)dF(\theta) \leq 0$. As such, if $\underline{\theta}^{\tilde{b}} \geq \tilde{\theta}^b$

$$\begin{aligned}
(V^*(b) - V^*(\tilde{b})) - (V^{\text{Pub}}(b) - V^{\text{Pub}}(\tilde{b})) &= \int_{\underline{\theta}^b}^{\underline{\theta}^{\tilde{b}}} v(1,\theta)dF(\theta) - \int_{\tilde{\theta}^b}^{\tilde{\theta}^{\tilde{b}}} v(1,\theta)dF(\theta) \\
&= \int_{\underline{\theta}^b}^{\underline{\theta}^{\tilde{b}}} -\underline{u}(\theta)\frac{-v(1,\theta)}{\underline{u}(\theta)}dF(\theta) - \left(\int_{\tilde{\theta}^b}^{\tilde{\theta}^{\tilde{b}}} -\underline{u}(\theta)\frac{-v(1,\theta)}{\underline{u}(\theta)}dF(\theta) \right) \\
&\leq \frac{-v(1,\underline{\theta}^{\tilde{b}})}{\underline{u}(\underline{\theta}^{\tilde{b}})} \int_{\underline{\theta}^b}^{\underline{\theta}^{\tilde{b}}} -\underline{u}(\theta)dF(\theta) - \left(\frac{-v(1,\tilde{\theta}^b)}{\underline{u}(\tilde{\theta}^b)} \int_{\tilde{\theta}^b}^{\tilde{\theta}^{\tilde{b}}} -\underline{u}(\theta)dF(\theta) \right) \\
&\leq \frac{-v(1,\tilde{\theta}^b)}{\underline{u}(\tilde{\theta}^b)} \left(-\int_{\underline{\theta}^b}^{\underline{\theta}^{\tilde{b}}} \underline{u}(\theta)dF(\theta) + \int_{\tilde{\theta}^b}^{\tilde{\theta}^{\tilde{b}}} \underline{u}(\theta)dF(\theta) \right) \leq 0
\end{aligned}$$

whereas if $\underline{\theta}^{\tilde{b}} \leq \tilde{\theta}^b$,

$$\begin{aligned}
(V^*(b) - V^*(\tilde{b})) - (V^{\text{Pub}}(b) - V^{\text{Pub}}(\tilde{b})) &= (V^*(b) - V^{\text{Pub}}(b)) - (V^*(\tilde{b}) - V^{\text{Pub}}(\tilde{b})) \\
&= \int_{\underline{\theta}^b}^{\tilde{\theta}^b} v(1,\theta)dF(\theta) - \int_{\underline{\theta}^{\tilde{b}}}^{\tilde{\theta}^{\tilde{b}}} v(1,\theta)dF(\theta)
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\underline{\theta}^b}^{\tilde{\theta}^b} -\underline{u}(\theta)\frac{-v(1,\theta)}{\underline{u}(\theta)} dF(\theta) - \left(\int_{\underline{\theta}^{\tilde{b}}}^{\tilde{\theta}^{\tilde{b}}} -\underline{u}(\theta)\frac{-v(1,\theta)}{\underline{u}(\theta)} dF(\theta) \right) \\
&\leq \frac{-v(1,\tilde{\theta}^b)}{\underline{u}(\underline{\theta}^{\tilde{b}})} \int_{\underline{\theta}^b}^{\tilde{\theta}^b} -\underline{u}(\theta) dF(\theta) - \left(\frac{-v(1,\underline{\theta}^{\tilde{b}})}{\underline{u}(\tilde{\theta}^b)} \int_{\underline{\theta}^{\tilde{b}}}^{\tilde{\theta}^{\tilde{b}}} -\underline{u}(\theta) dF(\theta) \right) \\
&\leq \frac{-v(1,\tilde{\theta}^b)}{\underline{u}(\tilde{\theta}^b)} \left(- \int_{\underline{\theta}^b}^{\tilde{\theta}^b} \underline{u}(\theta) dF(\theta) + \int_{\underline{\theta}^{\tilde{b}}}^{\tilde{\theta}^{\tilde{b}}} \underline{u}(\theta) dF(\theta) \right) \\
&= \frac{-v(1,\tilde{\theta}^b)}{\underline{u}(\tilde{\theta}^b)} \left(- \int_{\underline{\theta}^b}^{\underline{\theta}^{\tilde{b}}} \underline{u}(\theta) dF(\theta) + \int_{\tilde{\theta}^b}^{\tilde{\theta}^{\tilde{b}}} \underline{u}(\theta) dF(\theta) \right) \leq 0
\end{aligned}
$$

This proves the claim. □