Beliefs from Cues

John J. Conlon Spencer Y. Kwon

January 27, 2025*

Abstract

We develop a model of belief formation in which normatively irrelevant cues influence which states come to mind. Beliefs depend on two key factors: representativeness, the similarity of a cue to high-utility states, and resonance, the cue's ability to evoke many states. We show how to derive testable predictions from the model from subjective similarity judgments, a method we showcase in a controlled experiment in which participants estimate probabilities in a task involving uninformative cues. Beliefs systematically shift with cues' (separately measured) resonance and representativeness, consistent with the model's predictions. Finally, we explore economic applications of our framework, showing how cue-based beliefs can underlie (and make novel predictions about) seemingly disparate behavioral phenomena, such as experience effects, projection bias, present bias, uninformative advertising, and underdiversification.

*Conlon: Carnegie Mellon University, jconlon@andrew.cmu.edu. Kwon: Brown University, spencer_kwon@brown.edu. We thank Francesca Bastianello, Pedro Bordalo, Nicola Gennaioli, Botond Koszegi, Rafael La Porta, George Loewenstein, Muriel Niederle, Ryan Oprea, Chris Roth, Josh Schwartzstein, Andrei Shleifer, Adi Sunderam, Ansgar Walther, and seminar audiences at the Bravo Center Conference for Behavioral Economics, SITE Experimental, the Bocconi Cognitive Foundations of Finance Conference, and the IZA Beliefs Workshop for helpful comments and feedback.

1 Introduction

Standard models in economics—even those incorporating behavioral frictions—tend to assume that agents' beliefs are a function only of objectively relevant pieces of information, such as their priors and the likelihood of signals. At the same time, there is a widespread understanding that this picture must be incomplete: elicited expectations seem overly sensitive to respondents' local context, advertisements meant to persuade consumers include many seemingly irrelevant features, and experimental economists worry endlessly that their survey design or task framing will prime individuals to give different beliefs than they "normally" would.[1] In other words, beliefs appear sensitive to normatively irrelevant *cues*. While this (sometimes implicit) knowledge is ubiquitous, the role of cues in shaping expectations has thus far largely resisted formal economic analysis.[2] How can we make progress on incorporating cues into economic models? What data are required to discipline such a theory and yield testable predictions from it, and what implications do cue-based beliefs have for economic choices?

In this paper, we describe and experimentally test a model of beliefs from cues. Our agent decides whether to take an action—for example, purchasing a product or investing in a particular portfolio—whose value depends on the state of the world. A standard rational agent would simply integrate over all possible states to form her belief about the action's expected utility. Our agent, instead, is more likely to think of states that are *similar* to the context she is forming her belief in. We model the context as a collection of cues—endogenous or exogenous—that bring certain states to mind more than others. The strength of this association between cues and states depends on their similarity to each other.

When the context consists of a single cue, we show that belief distortions (i.e., the difference between the agent's belief and the truth) depend on the correlation between how similar the cue is to various states and the utility the action yields in those states. Borrowing from Kahneman & Tversky (1972), we call this the cue's *representativeness* (with respect to high-utility states).[3] When the context contains multiple cues, cue-based dis-

[1] See Taubinsky et al. (2024) for evidence on expectations being excessively sensitive to local contexts. The fact that many advertisements appear to lack informational content has long been documented (Stern & Resnik 1991) and studied by economists (e.g, Milgrom & Roberts 1986, Becker & Murphy 1993). For reviews of experimental/survey design, see Haaland et al. (2023) and Stantcheva (2022), each of which discusses the hazards that framing poses for researchers.

[2] Empirical tests of these forces instead rely heavily on researchers' intuitions about how particular cues might operate, with resulting disagreements and inconsistencies in results across experiments. See Cohn & Maréchal (2016) and Chivers (2019) for helpful discussions/reviews.

[3] Kahneman & Tversky (1972) describe the representativeness of an instance as in part depending on how similar it is to the population it comes from. Our notion of the representativeness of a cue (with

tortions are a convex combination of the representativeness of each component cue. The weight a particular cue receives depends on its *resonance*: that is, how many states are similar to it and therefore come to mind when it is part of the agent's context. Intuitively, even a highly representative cue may fail to move beliefs if it does not resonate: that is, though the states similar to a cue may be ones with extreme utility values, it may fail to move beliefs appreciably if these states are few in number and therefore outweighed by those of other cues.

We then show how to derive simple, testable comparative statics from the model. Predicting the effect that a cue will have on beliefs requires estimates of its resonance and representativeness, which ultimately depend on the underlying similarity between states. Though economists do not typically collect data on similarity, cognitive psychologists have long elicited and analyzed data on subjective similarity judgments to map out associations (e.g., Tversky & Gati 1982, Ashby & Perrin 1988, Nosofsky 1992). With data on such judgments, one can compute the resonance and representativeness of any given cue and then test whether the effect of cues on beliefs depends in the predicted way on these similarity-derived quantities.

We showcase this method, and thereby test the model, in a simple controlled experiment. Participants are tasked with forming beliefs about a simple data-generating process, framed as a "game" that can either result in a win or a loss. Each game consists of seven rounds that either add or subtract points from its total score. The number of contributed points each round is uniformly and independently drawn from the set $\{+1, +2, +3, +4, -5, -6\}$. A game results in a win if the sum of contributed points across all seven rounds is strictly positive. Participants must guessing the share of games that result in a win vs a loss. The large number of states (6^7) precludes participants from simply enumerating all the wins and losses, raising the possibility that cues may influence which states come to mind and therefore what beliefs they form.

To test this possibility, participants are randomized to either form beliefs absent any additional cue, or they are shown one or more cues framed as example games that they can watch while they form their beliefs. These example games are objectively uninformative: the correct answer to the question participants are forming beliefs about already logically follows the description of the rules of the game without any cue. Thus, the rational benchmark—and indeed the benchmark for any agent who responds only to objectively informative signals (e.g, even one who under- or overreacts to data)—is that these cues should have no effect on beliefs. Furthermore, participants are explicitly told that examples are randomly assigned and should not have any effect their beliefs. After

respect to a utility function) depends on how similar the cue is to states that yield high utility.

eliciting beliefs, participants are asked to make pairwise similarity judgments between example games. We then use these data to estimate the resonance and representativeness of cues.

We find that these cues have large effects on beliefs, and that these effects are systematically related to their (separately measured) resonance and representativeness. Moving from a cue that is representative of losses to one representative of wins increases beliefs about the share of wins from 45.6% to 62.3% (87% of a standard deviation, $p < 0.01$), with control-group beliefs in between these two values (and significantly different from both, $p < 0.05$ for both comparisons). We then estimate the model using average beliefs across 136 randomly generated cues. We find that our model explains over 57% of the maximum explainable variation in the data, despite having only two free parameters (resonance and representativeness of the underlying latent context).

Finally, we provide evidence on the effect of providing multiple cues. Recall that beliefs are a convex combination of the representativeness of the cues the agent faces. An implication of this result is that while providing multiple (say) positive cues can continue to increase beliefs beyond the effect of a single cue, there are sharp limits to the belief changes that can be achieved through cues alone. In particular, the maximum representativeness of a set of cues represents an upper bound on the belief distortions that set can induce. To test this prediction, we provide some participants with either zero (our control group), one, five, or ten cues, randomly drawn from a set of what we call our Baseline Win and Loss cues. We find that, relative to providing only one cue, providing five Baseline Win/Loss cues further moves beliefs toward those cues' representativeness. However, providing ten such cues has no additional effect relative to five, with beliefs stabilizing close to their representativeness.

Having provided experimental evidence for the model, we then show that beliefs from cues may help to unify and shed additional light on seemingly disparate behavioral anomalies. We start by embedding our theory into an otherwise standard model of intertemporal choice under uncertainty, where an agent decides how much to consume today versus save for the future. The agent's ability to imagine future states depends on whether they are cued by her context, which we assume includes the states she has already lived through (including the current state). We derive a similarity-augmented Euler equation that specifies how two factors distort the agent's savings decisions away from the rational benchmark: what we call differential cueing and cognitive discounting.

First, the agent's savings decision depends on whether her context (past and current states) differentially cues high marginal-utility future states. This leads the agent to exhibit both *experience effects* (Malmendier & Nagel 2011)—acting as though the future will

resemble states she has personally lived through—and *projection bias* (Loewenstein et al. 2003)—acting as though the future will resemble today. If she has lived through high marginal-utility states (e.g., unemployment or the Great Depression), or if she is currently in such a state, she will save more and consume less (than the rational benchmark), even if these cues are not objectively informative about the distribution of future states. Critically, these effects operate through, and therefore are moderated by, similarity: an agent whose home was burglarized will later appear more risk-averse over theft-related expenses (e.g., renter's insurance) but *less* risk averse in dissimilar domains (e.g., flood insurance). Bringing one state to mind crowds out others, and so the agent will exhibit inconsistent and sometimes paradoxical risk attitudes.

Second, the extent to which the future—and therefore the need to save—comes to mind at all depends on how much the present cues it, a force we call cognitive discounting. Thus an agent will look more patient when her present circumstances are similar to many future states but more imprudent in unusual circumstances: she will scrimp at the grocery store but splurge while on a vacation. This same force could operate over the life cycle: an agent will become more patient over time the longer she has lived in situations that resemble the (likely) future. Thus a worker who has previously bounced between jobs or locations will slowly begin to save more once she lands a steady career in a city she likes (the opposite of what a standard model would predict given the reduction in uncertainty).

Further, on the intuitive assumption that the distribution of states is stationary, the agent exhibits dynamically-inconsistent *present bias* (Laibson 1997, O'Donoghue & Rabin 1999): her discount factor is especially severe over the immediate future but levels off over longer horizons. Similarity drives this result. The "dropoff" in similarity to the present (and therefore in ease of imagination) between today and tomorrow can be severe, as any unusual present circumstances will likely disappear. But the dropoff in similarity to the present between 365 days from now and 366 days from now will likely be negligible (and, in the limit, zero). Thus our agent is especially impatient regarding benefits that can accrue in the near future, but will trade off benefits between more distant dates (asymptotically) rationally.

So far, we have considered a case where an agent faced exogenously given cues. Our experiment showed however that externally provided cues have large effects on beliefs, even in cases where agents are explicitly told they are normatively irrelevant. This raises the question of how a persuader, armed with the ability to cue strategically chosen states, can manipulate the agent to her advantage. To study this, we model a persuader that designs a product along with a cue to boost the agent's willingness to pay for it. For

example, a financial advisor might both construct a portfolio of assets and then try to "sell" her client on it by cuing her with its upside. Or a firm might produce a consumer good—whose overall usefulness may depend on how often the customer finds herself in certain situations (e.g., sunny weather for a convertible car)—as well as an advertisement to help its customers imagine how they will use it. We then ask what cue the persuader chooses as well as how the ability to provide cues influences the product she designs.

Our main result highlights a novel tradeoff between riskiness and *persuade-ability*. Consider a financial advisor designing a portfolio for a risk-averse client. A perfectly safe asset yields the same return in all states, but for precisely this reason it is impervious to cueing: there is no cue the advisor could provide that would boost the agent's beliefs about its expected return. In contrast, a risky portfolio (one, say, heavily invested in a particular industry) raises the possibility of persuasion through cueing ("The tech sector might take off!"). The client therefore ends up with an under-diversified portfolio, because it maximizes the fees the advisor can extract. In a consumer-product setting, this same force leads to greater *ex post* regret than in the rational benchmark. The firm designs more "specialized" products—like warranties or time-share vacation homes—with particular, high-utility use cases, which the firm cues in its advertisements to customers. Buyers in turn end up in these high-utility situations less than it felt like they would when they purchased the product. These analyses show how cue-based persuasion can affect more tangible aspects of the economy like portfolio compositions and product design.

This paper contributes first to a literature on decision-making under complexity. Recent work considers how beliefs and actions are distorted when agents cannot retrieve all relevant information, consider all relevant hypotheses or features, and/or noiselessly implement decision procedures (e.g., Bordalo, Gennaioli, et al. 2024a, Woodford 2020, Caplin et al. 2019, Gabaix 2019, Enke & Graeber 2023, Bordalo et al. 2023a, Graeber et al. 2023, Ba et al. 2023, Bohren et al. 2024, Oprea 2024, Conlon 2024, Augenblick et al. 2024). Our paper also speaks to the related literature on failures of contingent reasoning (see Niederle & Vespa 2023 for a review), where agents fail to fully or accurately consider states in which their actions are especially relevant (e.g., Esponda & Vespa 2014, Martínez-Marquina et al. 2019, Moser 2019).[4] Some of this literature studies cues, though typically only regarding their ability to bring to mind past *memories* (Bordalo et al. 2023b,

[4]An interesting difference between our framework and the literature on failures of contingent reasoning is that these papers often focus on settings with few (e.g., only two) relevant states; we speculate that one force driving some failures of contingent reasoning in these studies may be the difficulty of correctly or fully simulating a given state. For example, participants in Esponda & Vespa (2014)'s experiment appear not to realize absent prompting that there is *no* state where one action out-earns the alternative action. We abstract away from this consideration in our study and instead focus on the question of *which* states people simulate, rather than their ability to accurately simulate a given state conditional on attempting to do so.

Bordalo, Burro, et al. 2024b, Enke et al. 2024, Bordalo, Gennaioli, et al. 2024b). Our work shows that a similar framework can be productively applied to study which *unrealized* future states come to mind when agents form beliefs.

Our work also complements a growing literature on changing mental models (Eliaz & Spiegler 2020, Schwartzstein & Sunderam 2021, Barron & Fries 2023, Charles & Kendall 2024). Existing theories tend to assume that, of the models people are exposed to, they adopt only one (e.g., the best-fitting model or the one that delivers the highest anticipatorty utility). Our focus instead is on associations: our agent equally "counts" all states that she thinks of, but which states come to mind depends on the similarity between them and the cues she faces.

Next we contribute to a literature on simulation in economics and psychology. The role of simulation (the process of imagining future states/scenarios) in forming beliefs has long been recognized and studied in psychology (e.g., Kahneman & Tversky 1981, Dougherty et al. 1997, Gershman et al. 2017, Chater et al. 2020). It has received less attention in economics, with some important exceptions. For example, Bordalo, Burro, et al. (2024b) study how memories of past experiences can serve as cues to aid or inhibit simulation of future risks. Becker & Mulligan (1997) models an agent who can invest in her ability to appreciate the future, while Gabaix & Laibson (2022) study how an agent who (realizes she) can only approximate future utility flows noisily will often seem to act impatient or present biased.

Lastly, as alluded to above, our model closely resembles theories of memory retrieval (Kahana 2012, Bordalo et al. 2023b), reflecting work in psychology finding overlap between the brain processes underlying simulation and memory (Schacter et al. 2007, Gilbert & Wilson 2007, Schacter et al. 2012, Mullally & Maguire 2014, Benoit & Schacter 2015). Our focus on the importance of similarity between states in facilitating simulation echoes other work in economics studying similarity as a underlying driver of many cognitive processes, such as memory retrieval (Mullainathan 2002, Enke et al. 2021, Charles 2022b,a, Bordalo et al. 2023b, Graeber et al. 2023, Link et al. 2023, Bohren et al. 2024, Colonnelli et al. 2024), categorization (Mullainathan et al. 2008, Bordalo, Gennaioli, et al. 2024a, Evers et al. 2021), reinforcement learning (Barberis & Jin 2023), and reasoning by analogy (Gilboa & Schmeidler 1995).

2 Model

In this section, we first describe the basic setup of the model, which formalizes how an agent forms beliefs by sampling states of the world. States come to mind more readily

when they are more similar to the cues the agent faces while forming beliefs, leading to belief distortions. We show that two properties of cues—summarizing whether they bring many states to mind and the extent to which those states favor a particular hypothesis—determine their effect on beliefs. These lead to testable comparative statics based on the underlying similarity structure of the state space, which can be proxied for by eliciting similarity judgments. We then conclude this section by discussing the limits of cue-based belief distortions.

2.1 Setup

Objective environment Let Ω be the set of possible states of the world, with the likelihood (or frequency) of $\omega \in \Omega$ given by $\pi(\omega)$. The agent is considering an action whose payoff $\psi(\omega)$ is state-dependent. Under the frictionless rational benchmark, the expected utility of the action is then given by $\Psi_{rat} \equiv \sum_{\omega \in \Omega} \pi(\omega)\psi(\omega)$.

Similarity The key idea that our model formalizes is that, unlike a standard rational agent, an agent may struggle to consider all possible states, and the context in which she forms her beliefs may influence which states come to mind. In particular, we assume that states are more likely to come to mind the more similar they are to the context the agent faces (Bordalo et al., 2023b). This implies that normatively irrelevant elements of her context—which we call "cues"—can distort her beliefs by bringing similar things to top of mind.

Formally, we assume an exogenous similarity function $S : \Omega \times \Omega \mapsto [0,1]$ on the state space, where $S(\omega,\omega) = 1$ for all ω. $S(\omega_1,\omega_2)$ captures the degree to which ω_1 is similar to (or "associated with") ω_2. We can then generalize S to apply to weighted subsets of Ω by averaging element-wise similarity. Formally, let A and B be weighted subsets with support Ω, with $w_A(\omega), w_B(\omega)$ be the weight of ω in A and B. Then, the similarity between A and B is given by:

$$S(A,B) = \frac{1}{|A| \cdot |B|} \sum_{\omega,\omega' \in \Omega} w_A(\omega) w_B(\omega') \cdot S(\omega,\omega'), \tag{1}$$

where $|A| = \sum_\omega w_A(\omega)$ and $|B| = \sum_\omega w_B(\omega)$.

Context and cued beliefs We focus on cues that can be modeled as elements of Ω. Formally, we define the local context $C = \{\omega, w_C(\omega)\}$ as a weighted subset of Ω, with $w_C(\omega)$ capturing the weight of ω in C. We assume the agent simulates utility Ψ_{rat} by imagining

(or "simulating") possible states. Drawing from past work (Bordalo et al., 2023b), we model this as an associative sampling process. In particular, we assume the agent has an unbiased "mental database"—she is aware, in a sense, of all possible states ω and their objective likelihood $\pi(\omega)$—but the sampling process that allows her to access these states is distorted by the context C. In particular, we assume the probability she samples a state ω is proportional to its similarity to the context $S(\omega, C)$. In the large sample limit, the agent's simulated distribution is then proportional to:

$$\pi_s(\omega|C) \propto \pi(\omega) \cdot S(\omega, C), \tag{2}$$

where $S(\omega, C)$ is given by Equation 1. We assume that the agent is naive about the context distortion: her beliefs are thus given by the simulated distribution $\pi_s(\omega|C)$. Equation 2 highlights that the beliefs cued by the context can be simply viewed as the true distribution π with context-dependent distortion weights $S(\omega, C)$. When sampling is unbiased ($S(\omega, C) = 1$), one recovers the rational benchmark.

Interpretation We view our model as providing a tractable way of incorporating associative (i.e., similarity-based) cues into a model of belief formation. Naturally, many assumptions are simplifications. For example, the similarity function in our model is exogenous, but other work microfounds and endogenizes similarity by considering how many salient features two items have in common (Tversky, 1977; Bordalo et al., 2023a). We also model cues simply as subsets of the state space, which is restrictive: for example, our model will fail to incorporate effects of describing a state in different terms or of cues that differ greatly in type from states of the world (e.g., an agent's current emotions). But our framework provides a natural starting place for analyzing the effect of many common cues: for example, the agent's current situation, her experience of past states, depictions or descriptions of scenarios (e.g., in advertisements), and so on.

Finally, though for simplicity we describe our agent as making a choice under uncertainty—where she is evaluating the probability of different states of the world—our framework can be applied more generally. The crucial ingredients are that there be 1) a large set of objects about which agents are forming beliefs, and 2) an *associative structure* over these objects. Though risky choice under unknown states of the world is a natural example of a setting with these two inputs, there are others that lack any "true" uncertainty. For example, an agent may be forming beliefs about her own past experiences, where her recall of memories is partly shaped by their associations (Bordalo et al., 2023b). Or an agent may be trying to evaluate how often she will find herself in various

situations in the future, about which she may in principle already know enough to decide with certainty. All of the analysis to follow could equally apply to a persuader trying to convince an agent in settings like these where there is no "true" uncertainty.

2.2 Characterizing cued beliefs

Characterizing beliefs To build intuition, we first focus on the case where $\psi(\omega) = I(\omega \in H)$, $H \subset \Omega$. In that case, $\Psi_{rat} = \pi(H)$: the agent is assessing the likelihood of a hypothesis H, which is a subset of the state space. Equation 2 implies the following lemma regarding how a given context $C = \{\omega_c, w(\omega_c)\}$ distorts the assessment of $\pi(H)$.

Lemma 1. *Let $\mathcal{W}(\omega_c) = E[S(\omega, \omega_c)]$ be the average similarity of ω_c to $\omega \in \Omega$ and $\mathcal{D}^H(\omega_c) = \frac{\sum_{\omega \in H} S(\omega,\omega_c)\pi(\omega)}{\sum_{\omega \in \Omega} S(\omega,\omega_c)\pi(\omega)} - \pi(W)$. The agent's beliefs given context $C = \{\omega_c, w(\omega_c)\}$ are given by:*

$$\underbrace{\pi_s(H|C) - \pi(H)}_{\textit{Context distortion}} = \frac{\sum_{\omega_c \in C} w(\omega_c) \cdot \mathcal{W}(\omega_c) \cdot \mathcal{D}^H(\omega_c)}{\sum_{\omega_c \in C} w(\omega_c) \cdot \mathcal{W}(\omega_c)}. \tag{3}$$

Lemma 1 shows that the effect of $\omega_c \in C$ on the assessment of $\pi(W)$ is characterized by two summary statistics based on similarity. The first is $\mathcal{W}(\omega_c) = E[S(\omega, \omega_c)]$, which we call the *resonance* of ω_c, which is the average similarity of ω_c to other states in Ω. Because cues bring states to mind in proportion to their similarity, the resonance of ω_c captures the extent to which ω_c brings many states to mind. The second quantity, $\mathcal{D}^H(\omega_c)$, is what we call the *representativeness* of ω_c with respect to hypothesis H. $\mathcal{D}^H(\omega_c)$ intuitively captures whether the states that ω_c bring to mind tend to be consistent with hypothesis H. Equation 3 implies that the effect of context C on the agent's beliefs about H is a convex combination of the representativeness of each of its constituent cues, with weights proportional to their resonance.

Figure 1 visualizes the effect of a context containing two cues on the agent's beliefs. We consider the case in which $\Omega = \mathbb{R}$, with the true distribution π given by the normal distribution $\mathcal{N}(0, 1)$. We endow Ω with the similarity function $S(\omega, \omega') = \exp\left(-\frac{\kappa}{2}(\omega - \omega')^2\right)$. Suppose that the agent is assessing the likelihood of $H = \{\omega \geq 0\}$, where the context consists of two equally weighted cues: $C = \{-2, 1\}$. The red and blue curves in Figure 1 plot respectively $S(\omega, 1) \cdot \pi(\omega)$ and $S(\omega, -2) \cdot \pi(\omega)$, the beliefs that each cue would induce absent the other. The resonance of each cue is given by the total mass under each curve, while the representativeness of each cue with respect to H is given by the ratio between the darker and lighter areas for each curve. As the figure illustrates, $\omega_c = 1$ both has higher resonance and representativeness (with respect to H) than $\omega_c = -2$. The resulting

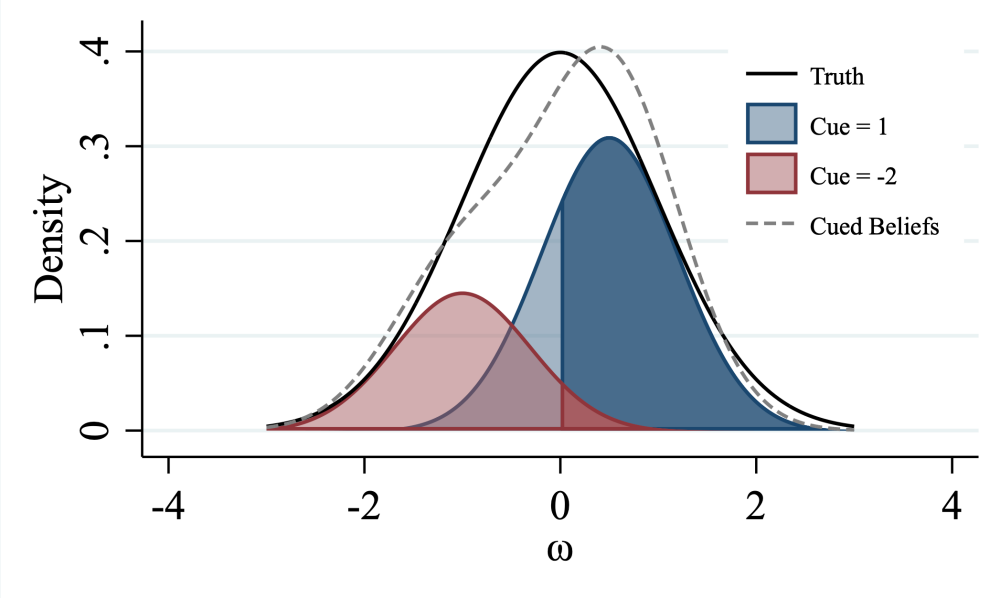


Figure 1: Context-distorted beliefs

context-distorted beliefs are shown by the dashed curve.

Proposition 1 generalizes the expression in Lemma 1 to any utility function $\psi(\omega)$:

Proposition 1. *The agent's beliefs $E_s[\psi(\omega)|C]$ in context C are:*

$$\overbrace{E_s[\psi(\omega)|C] - E[\psi(\omega)]}^{\textit{Effect of cues}} = \frac{\sum_{\omega_c \in C} w(\omega_c) \cdot \mathcal{W}(\omega_c) \cdot \mathcal{D}^{\psi}(\omega_c)}{\sum_{\omega_c \in C} w(\omega_c) \cdot \mathcal{W}(\omega_c)}, \tag{4}$$

where $\mathcal{D}^{\psi}(\omega_c) = \frac{Cov[S(\omega,\omega_c),\psi(\omega)]}{E[S(\omega,\omega_c)]}$ is the representativeness of ω_c with respect to ψ.

Note that the only difference between the expressions in Lemma 1 and Proposition 1 are in the definition of representativeness. Representativeness in Proposition 1 is defined as the covariance between the similarity to ω_c and the value of ψ. Thus, it captures the representativeness of ω_c with respect to high-utility states. The function is equivalent to the previous notion of representativeness when $\psi(\omega) = I(\omega \in H)$.

2.3 Limits to cues

Aggregation limits While our framework predicts that beliefs are malleable to cues and irrelevant changes in the context, our theory also predicts limits to such instability. In particular, context-driven belief distortion is constrained by the maximal representa-

tiveness of a cue:

$$E_s[\psi|C] - \Psi_{rat} \leq \max_{\omega \in C} \mathcal{D}^{\psi}(\omega). \tag{5}$$

Equation 5 holds even in the limit as $|C| \mapsto \infty$, illustrating a stark *failure of aggregation*. The intuition for this result is simple: cues move beliefs by shifting which states are more available and top of mind. When there are multiple cues, cues *interfere* with each other – making one state more top of mind crowds out other states. Furthermore, the limits to belief distortion are governed not by objective probability, but the underlying associative structure, a point we turn to next.

When are cues effective? To see further implications of equation 5, we impose the following similarity structure on Ω (Tversky, 1977). An element of $\omega \in \Omega$ is defined by a finite set of K features $\omega = (f_1(\omega), f_2(\omega), ...f_K(\omega))$, with $S(\omega, \omega') = \prod_{k=1}^{K} \delta_k^{I(f_k(\omega) \neq f_k(\omega'))}$. The parameter δ_k captures whether feature f_k is distinctive or not: the more distinctive f_k, the lower the δ_k – states that disagree on the distinctive feature are judged as less similar. Suppose an agent is assessing the likelihood of $H = \{f_k(\omega) = \bar{f}\}$, and assume for simplicity that $\delta_{k'} = 1$ for $k' \neq k$. Then, one obtains the following upper bound on the maximal representativeness of a cue with respect to H, which is also then an upper bound to context-distortion:

$$\pi_s(H|C) - \pi(H) \leq \max_{\omega} \mathcal{D}^{H}(\omega) \leq \frac{1 - \delta_k}{\delta_k + \pi(H)/(1 - \pi(H))} \cdot \pi(H). \tag{6}$$

As equation 6 immediately implies, the scope of increasing the subjective assessment of $\pi(H)$ through cues is greater for more distinctive hypotheses (lower δ_k). If H is not a very distinctive hypothesis—elements within H are not much more similar to each other than to elements in $\Omega - H$—it is difficult to find a cue that strongly boosts H and interferes with $\Omega - H$. In contrast, if the hypothesis is highly distinctive, cuing can have a large effect on assessments. The problem of which random variables $\psi(\omega)$ are most malleable to cues is essential to the theory of endogenous cues, which we revisit in Section 4.2.

2.4 Testable predictions

Though the theory described above is simple, it contains two objects—the agent's context C and the similarity function S—that we have intentionally left quite general. The context might contain many cues: the agent's current situation, past experiences, descriptions/depictions of future scenarios, or any items in her environment that bring particular possibilities to mind. And many features of these cues might in turn influence their

similarity to states of the world and therefore their effect on the agent's beliefs. One might worry that such generality leaves the model without testable predictions: any pattern of beliefs could in principle be rationalized by postulating an appropriate cue and similarity function. What, then, is the empirical content of the theory?

Fortunately, two observations allow us to formulate testable predictions. The first is that, while the similarity between cues and states might be multifaceted, it is measurable. Decades of research in cognitive psychology has studied associations by analyzing subjective similarity assessments elicited from experimental participants (e.g., Nosofsky 1988, Tversky 1977), and recent work in economics has begun to adopt these insights (Bordalo et al. 2023a). Furthermore, one need not map out the entirety of the similarity function S; Proposition 1 makes clear that it is sufficient to estimate the resonance and representativeness of cues.

Second, while the context C might contain many cues—such that identifying them all and characterizing their resonance/representativeness would be infeasible—we can nonetheless derive comparative statics regarding the effect of a *marginal* cue. This is a familiar situation to economists: for example, we cannot observe the entire bundle of amenities that a job offers its employees, but we can derive predictions regarding the effect that a change in one of them (e.g., the wage) ought to have on worker behavior.

Formally, suppose we wish to predict the effect that adding a cue Q to the context C will have on the agent's beliefs. Equation 2 implies:

$$E_s[\psi(\omega)|C \cup Q] - E_s[\psi(\omega)|C] = \frac{\mathcal{W}(Q)}{\mathcal{W}_C + \mathcal{W}(Q)}(\mathcal{D}^{\psi}(Q) - D_C^{\psi}) \tag{7}$$

The above equation emphasizes that the effect of adding Q depends only on the relative resonance and representativeness of Q compared to the remaining context C. If Q has a higher representativeness than C, adding it to the context will boost beliefs, but the extent of this change will depend on how resonant Q is compared to C. Beliefs thus can be seen as a "tug-of-war" between Q and C. Each element of the context pulls beliefs toward its respective representativeness, with its resonance determining the strength of this pull.

Testing such predictions requires estimates of the resonance and representativeness each cue Q among a set of possible cues $\mathbb{Q}$. These can be computed using the Monte Carlo method. Suppose we elicit pairwise similarity judgments between Q and N random

samples $\omega_{1\leq k\leq n} \sim \pi(\omega)$. We can then compute for each $Q \in \mathbb{Q}$:

$$\widehat{\mathcal{W}}(Q) = \frac{1}{n}\sum_{k=1}^{n} S(\omega_k, Q) \qquad \widehat{\mathcal{D}^{\psi}(Q)} = \frac{\sum_{k=1}^{n} S(\omega_k, Q)\psi(\omega_k)}{\sum_{k=1}^{n} S(\omega_k, Q)} - \Psi_{rat}.$$

With such estimates of $\mathcal{W}(Q)$ and $\mathcal{D}^{\psi}(Q)$ in hand, we can then test simple comparative statics. Holding fixed $\mathcal{W}$, cues with a higher representativeness $\mathcal{D}^{\psi}$ should increase beliefs by more. Holding fixed representativeness $\mathcal{D}^{\psi}$, cues with a greater resonance should have a greater effect on beliefs.

To summarize, one can derive testable predictions about the effect that each of a set of cues $\mathcal{Q}$ will have on beliefs by eliciting appropriate similarity data. In the next section, we employ this procedure in a simple lab-experimental environment as a proof of concept. We show that objectively uninformative cues have large effects on beliefs that are predictable given separately elicited similarity judgments. We discuss at the end of the next section strategies for generating such predictions in less controlled environments.

3 Experiment

We test the model's predictions in a simple controlled environment. Experimental participants form beliefs about how often a "game" ends up being a "win" or "loss." We then test the impact of providing cues—highlighting particular outcomes of the game by having participants watch them occur—on participants' beliefs. We use similarity data elicited from participants to generate predictions for the effects of these cues, which we then test. Finally, we test the impact of providing multiple example games to explore the limits of persuasion through cues.

3.1 Design

The Environment. The experiment begins by describing to participants a simple data-generating process framed as a "game." Each game has seven rounds r, in each of which a number of points p_r is added to the "total score" of the game, where $p_r \in \{+1, +2, +3, +4, -5, -6\}$ with each outcome being equally likely and i.i.d. across rounds. To make this process intuitive for participants, it is described as rolling a six-sided die. The faces showing one through four are displayed in green, indicating that this number is added to the score, while five and six are shown in red, indicating subtraction. Participants are tasked with guessing the share of games that are "wins," meaning that they end with a total score (adding up across the seven rounds) that is strictly positive.

Note that this environment maps directly onto the model described above. In the experiment, the set of states $\Omega = \{+1, +2, +3, +4, -5, -6\}^7$ where as in Section 2 each state is equally likely. Further, these states are naturally grouped into a focal hypothesis $H = \{s : \sum_{r=1}^{7} s_r > 0\}$, the set of wins, with its complement being the set of losses. There are also a large number of states—6^7, or 279,936, of which 129,364 (46.2%) are wins—obviously precluding the strategy of simply enumerating all winning and losing possibilities. If participants instead form their beliefs relying on which outcomes come to mind easily, their beliefs might be susceptible to cues.

Cues. After learning the rules of the game, all participants are shown a graph to help them visualize how it works, as shown in the left panel of Figure 2. Those in the control group, who are shown no additional cue, must simply wait 20 seconds before a text box appears asking them the average number of games per 100 that are wins. Participants are aware that this is the question they will be asked during this waiting period.

The remainder of participants are shown between one and ten additional cues before the experiment elicits their beliefs. These cues are framed as example games that participants watch unfold by hovering their mouse over dots that sequentially depict results of each round of the game. See the right panel of Figure 2 for an image of a completed cue. Participants must wait 20 seconds after seeing each cue before either seeing an additional cue or before a text box appears eliciting their beliefs about the fraction of games that are wins.

Note that, for several reasons, these cues are objectively (and obviously) uninformative. First, the problem participants are solving does not involve any "true" uncertainty: the correct answer to the question they are answering is logically pinned down by the description of the data-generating process, which is fully specified even absent any example. Second, all outcomes are equally likely, so participants cannot infer that some outcomes are more likely from the fact that they were chosen to be shown as examples. And finally, we explicitly (and truthfully) tell all participants who receive a cue that while we are showing them, say, a win, we chose randomly whether they see winning or losing examples. Thus there is no objective signal to be extracted from the example game that participants see. Further, we explicitly state that "the example you are shown should not impact your guess", and participants must confirm their understanding of this fact in a comprehension question before continuing with the experiment.

We randomly choose example games to serve as cues. In particular, we randomly select two example games for ever possible final score between -40 and +27. This process yields 136 randomly generated cues. Participants are then sorted into the following

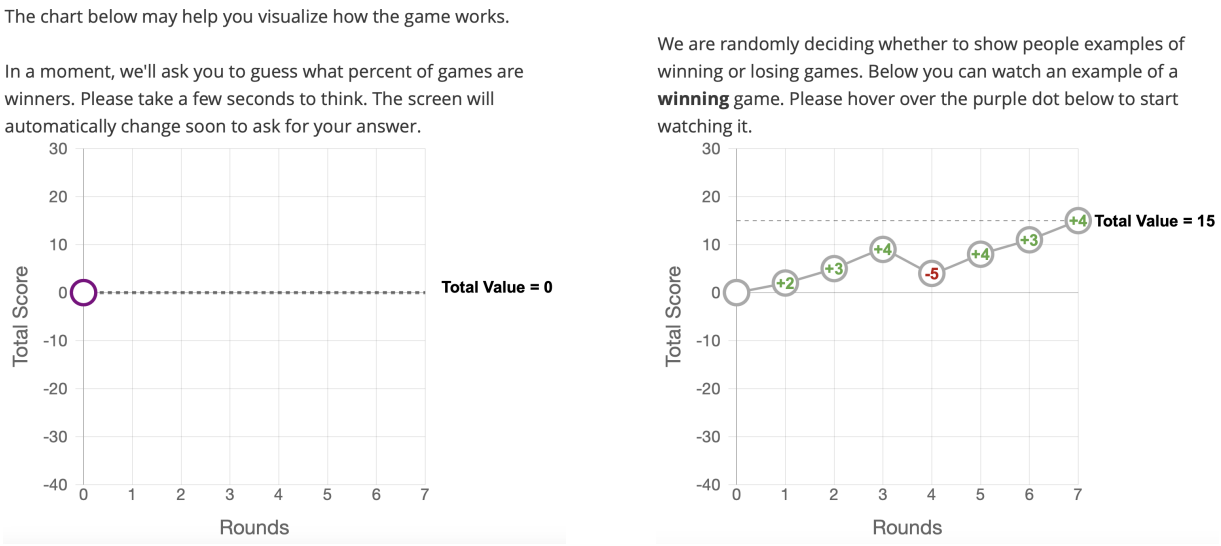


Figure 2: Belief elicitation with and without a cue

treatment groups:

- Control: Shown no additional cue, $N = 108$
- T1: Shown one cue randomly chosen from the 136 possible example games, $N = 1449$
- $T5_P$: Shown five distinct cues with final scores between +11 and +15, $N = 93$ (inclusive)
- $T5_N$: Shown five distinct cues with final scores between -15 and -11, $N = 117$ (inclusive)
- $T10_P$: Shown ten distinct cues with final scores between +11 and +15, $N = 125$ (inclusive)
- $T10_N$: Shown ten distinct cues with final scores between -15 and -11, $N = 103$ (inclusive)

Participants in the T5 and T10 treatment groups had to wait 10 and 5 seconds, respectively, between revealing the final outcome of one cue and proceeding to view the next cue. Note that these sample sizes ensure that approximately the same number of participants saw a single cue with a final score in the range +11 to +15 ($N = 110$) as saw five or 10 such cues in $T5_P$ and $T10_P$. Similarity, approximately the same number of participants saw a single cue with a final score between -15 and -11 ($N = 113$) as the number of participants seeing multiple such cues in $T5_N$ and $T10_N$. Henceforth, we refer to cues with final scores in these ranges as "Baseline Win" and "Baseline Loss" cues.

Similarity ratings. What effect should these cues have on participants' beliefs? The answer according to our framework depends on how similar they are to winning and losing outcomes, and therefore which states come to mind in response to them. We borrow from a literature in psychology studying similarity (e.g., Tversky & Gati 1982, Ashby & Perrin 1988, Nosofsky 1992) and elicit similarity judgments directly from participants. We then use those ratings to estimate the availability and diagnosticity of each cue we provide.

More precisely, all participants (regardless of treatment group) made 30 pairwise similarity judgments between outcomes of the game. Figure 3 shows what these elicitations looked like. Each pair of cues that participants compared included one cue from our set of 136 randomly generated cues as well as an example game randomly generated from the true data-generating process.

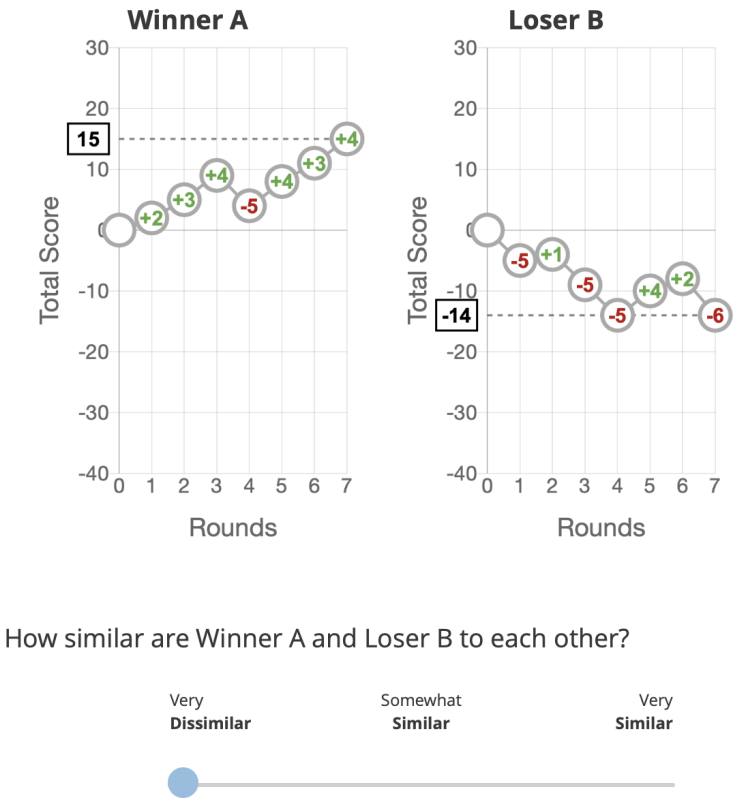


Figure 3: Similarity Elicitation

Following the discussing in Section 2.4, we then use these similarity data to estimate the resonance and representativeness (with respect to winning games) of each cue. In particular, the resonance of a cue is simply its average similarity to other states, which we compute by taking the sample average for each cue pooling across participants. The representativeness with respect to winning we similarity estimate by the Monte Carlo

method (see Section 2.4).

The left and middle panels of Figure 4 show the estimated resonance and representativeness of cues as their final score increases from -40 to 27. As is intuitive, the patterns we see are consistent with the final score playing a critical role in explaining participants' similarity judgments. First, cues with values closer to zero are more resonant, while those with more extreme values are least resonant: this reflects the fact that there are more outcomes with final scores close to zero and very few with extreme values. Second, the representativeness of cues with respect to winning increases (approximately) monotonically as their final score increases. Again this is intuitive: if participants in part focus on the final score as a determinant of similarity, than the games similar to a cue will tend to be winners (i.e., its representativeness will increase) if the cue has a higher final score.

Logistics. Participants were recruited from Prolific in October 2024. The median participant took 11 minutes to complete the experiment. Participants earned an additional $1.00 bonus if their belief about the fraction of winning games was within 5 percentage points of the truth.

3.2 Results

Reduced-form results. We begin by simply showing average beliefs across the treatments where participants were either shown no additional cue or one such cue. These averages are depicted in the right panel of Figure 4. We see that cues have substantial effects on beliefs, as we can reject that beliefs are equal across treatments at the $p < 0.001$ level. For example, moving from the Baseline Loss to Baseline Win treatment, beliefs increase from 45.6% to 62.3% ($p < 0.001$). This 16.7 percentage point increase represents 87% of the standard deviation of beliefs in the control group. Note that these effect appear despite all such participants being told that others were being shown example games of the opposite type (i.e., losses for Win treatment and wins for the Losses treatment). Thus, though participants are aware that the opposite examples exist, and are being shown to others in the experiment, actually "seeing" particular cues has substantial effects on beliefs.

How large are these effects? To answer this question, we use as a benchmark belief updating in response to objective signals from the literature. For example, in Möbius et al. (2022), participants receive a 75%-accurate binary signal about whether their IQ-test score was above the median. Beliefs about the chances of actually having an above-median performance are 48% of a standard deviation higher for those who receive a posi-

Table 1: Resonance and Representativeness Explain Cue Effectiveness

	Belief
Resonance × Representativeness	0.82***
	(0.27)
Resonance	0.22***
	(0.05)
Representativeness	0.07
	(0.10)
Observations	1,448

Notes: This table shows OLS regression estimates. The dependent variable is participants' beliefs about the share of games that are wins, restricting the sample to treatment group T1. The independent variables are the estimated resonance and representativeness of the cue they were shown, plus the interaction between these two quantities. Robust standard errors, clustered at the cue level, are presented in parentheses. *, **, and *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

tive vs a negative signal.[5] Thus, our uninformative cues have substantial effects on beliefs even compared to quite precise objectively informative signals.

The data presented so far constitute an existence result: we can reject the null hypothesis that cues, which are normatively irrelevant to rational agents, have no effect on beliefs. But do cues affect beliefs in the way theory from Section 2 predicts? Proposition 1 shows that we should expect an interaction effect between the representativeness and resonance of cues: more representative cues (w.r.t. winning) should boost beliefs about the share of winning games *particularly* when they have a high resonance. Table 1 shows that this is precisely what we find. In particular, we regress participants' beliefs on the resonance and representativeness of the cue those in T1 were shown, plus the interaction between these two quantities. The positive and significant ($p < 0.01$) coefficient on the interaction term reflects the fact that more resonant cues have a larger impact on beliefs, with the direction of that impact depending on their representativeness. We can visualize this interaction by comparing the three panels of Figure 4. Beliefs increase quickly with representativeness for cues close to zero but quickly tail off as the resonance of more extreme cues falls.

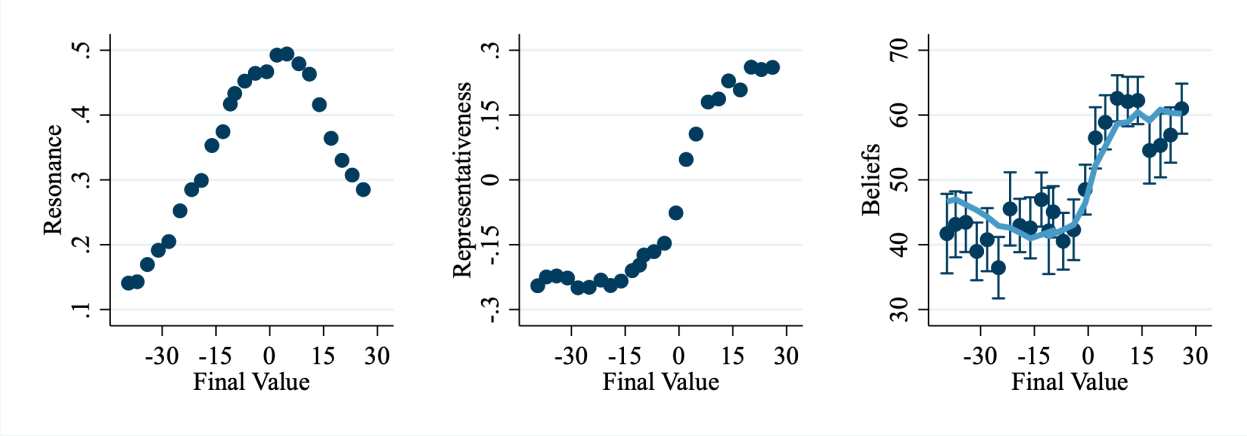
3.3 Structural estimation.

We can also use data from the Control and T1 treatments to structurally estimate the model. Note that, because the resonance and representativeness of cues are measurable, equation 7 has only two degrees of freedom: the resonance $\mathcal{W}_C$ and representativeness

[5]This statistic comes from our own calculation using Möbius et al. (2022)'s replication files.

D_C^{ψ} of the latent context. We use beliefs in the control group and for each cue in T1 to estimates these two parameters by the generalized method of moments. Our estimate for $D(d)$ is 0.53, which unsurprisingly is quite close to the the average belief in the control group (0.50). The estimate for $\frac{1}{\mathcal{W}_C}$ is 1.95, significantly greater than zero ($p < 0.01$), indicating (as Figure 4 already makes clear) that cues are effective in changing beliefs relative to the control group.[6] The right panel of Figure 4 plots the model-implied beliefs given these two parameters. We see a quite close quantitative fit between the model and the empirical treatment effects, despite the model being overidentified (138 moments to fit two parameters). More precisely, the R^2 between model-predicted and actual beliefs is 0.14, whereas a regression with cue-fixed effects (i.e., 136 indicators) yields an R^2 of 0.24 (adjusted R^2 = 0.17). Thus, our two-parameter model explains 58% of the maximum explainable variation in the data (Fudenberg et al., 2019).

Figure 4: Availability + Diagnosticity + Persuasiveness



Notes: This figure shows the resonance (left panel), representativeness (middle panel), and average beliefs (right panel) associated with cues in T1. The light blue curve shows the model-implied beliefs given our structural estimation.

3.4 Multiple cues.

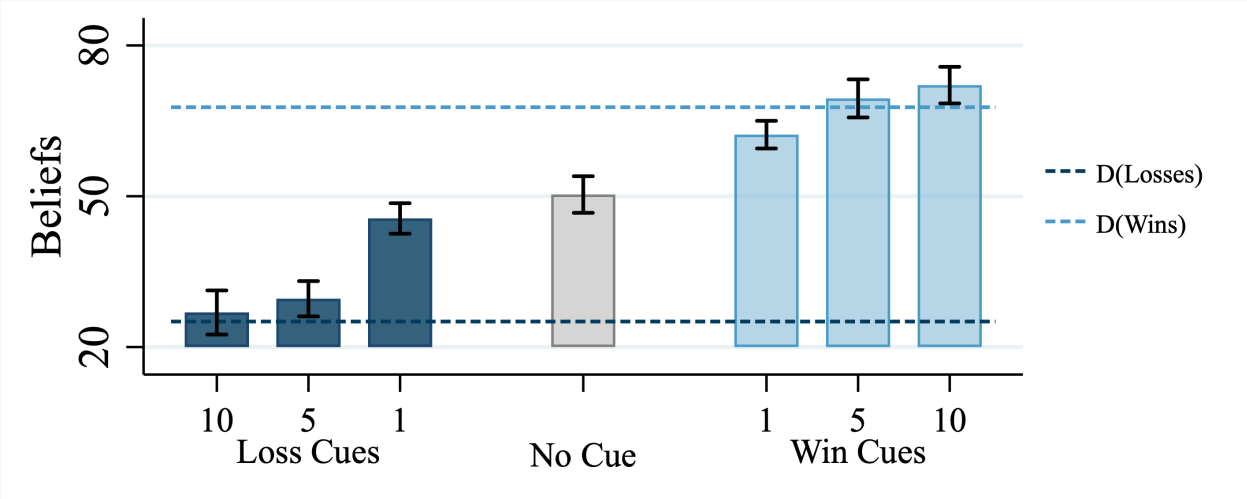
What effect does providing multiple additional cues have on beliefs? A natural supposition, which we explored in section 2.3, was that the total weight a set of cues gets in the agent's beliefs increases in their number. But even so, as we saw, persuasion through cues imposes limits on how much the agent's beliefs can change. In particular, though multiple cues might be more persuasive than only one, the maximum of their representativeness is an upper bound on the agent's beliefs. We now test this predictions.

[6] We estimate the reciprocal of $\mathcal{W}_C$ so that a coefficient of zero ($\mathcal{W}_C \to \infty$) corresponds to a case where cues have no effect on beliefs.

Recall that in our T1, T5, and T10 treatments, participants saw either one, five, or ten Baseline Win or Baseline Loss cues. This variation in the number of cues (holding set of cues fixed) allows us to test whether beliefs increase in the number of cues provided and, if so, whether beliefs remain constrained by their diagnosticity.

Figure 5 shows average beliefs, among participants who either saw a Baseline cue or were in the control group, split by the number of cues they were shown. The dotted lines in addition show the average representativeness of the Baseline Loss (dark blue) and Baseline Win (light blue) cues. Two patterns stand out, corresponding to the two predictions we are testing. First, additional cues are indeed more persuasive than only one. Beliefs increase from 62.3% to 69.5% ($p < 0.01$) as we move from one to five Baseline Win cues, and decrease from 45.6% to 29.6% ($p < 0.01$) when moving from one to five Baseline Loss cues. But, second, beliefs eventually level off around the level of the cues' representativeness as we move from five to ten such cues. We do not see a significant increase in beliefs moving from five to ten Baseline Win cues ($p = 0.32$) nor a significant decrease when moving from five to ten Baseline Loss cues ($p = 0.34$).

Figure 5: Multiple Cues and Aggregation Limits



To conclude, our experimental results show that cues can be quite persuasive even in an environment in which they are obviously objectively uninformative. Further, the magnitude of their effect on beliefs is modulated (and constrainted) by their representativeness and resonance, which we compute using separately elicited similarity judgments. This latter result points to the usefulness of collecting novel types of data to inform and discipline behavioral models.

3.5 Discussion: Similarity in Less Controlled Environments

One advantage of our experimental design is that the set of all possible states is known. This allows us to elicit similarity judgments between the cues we provide and (a random sample of) the full distribution of states. In less controlled environments, where the true state space is unknown, this method may not be available. How can similarity data be used to predict the effect of cues in such cases? While fully exploring this question is outside the scope of the present paper, two methods seem promising.

First, emerging machine-learning methods allow researchers to extract proxies for similarity from text data. For example, Sarkar (2024) uses text from financial news to extract measures of similarity between firms, which he shows predict excess returns. We view extending such methods to consider the similarity between cues (e.g., news events, pitches from financial advisors) and companies or past news events as a promising avenue for predicting their effects on investors, and similar methods could be applied to other domains. Second, recall that to predict the effect of cues, one needs only to estimate their representativeness and resonance. Eliciting proxies for these objects may be possible even when the full state space is unknown (and therefore the Monte Carlo method we employ in our experiment is unavailable). For example, one could ask participants directly whether states similar to a cue tend to adhere to a certain hypothesis (representativeness) or how likely are states similar to the cue (resonance).

4 Economic applications

Having provided experimental evidence on the model's main predictions, we now turn to investigating its economic implications. We first study exogenous cues. In particular, we focus on the case of intertermporal choice, where an agent's dynamic consumption decisions are distorted by cued past or current experiences. We show that our model naturally predicts experience effects, projection bias, and present bias, whose strength depends on the similarity between cues and future states. We then turn to endogenous cues, where a persuader jointly optimizes a product / service alongside a cue to boost the receiver's willingness to pay. When applied to financial persuasion, we document a novel tradeoff between risk and *persuade-ability*: cue-based persuasion most strongly benefits portfolios with risky / state-dependent returns.

4.1 Exogenous cues: application to intertemporal choice

Setup We consider the following simplified intertemporal setting. Let $\omega_t, \omega_{t+1}, ...$ follow an exogenous Markov process with *known* transition probability $\pi(\omega_{t+1}|\omega_t)$. At time t, the agent chooses a consumption plan $\vec{c_t} = (c_t, c_{t+1}, ...)$, subject to an intertemporal budget constraint $\sum_{h=0}^{\infty} \frac{c_{t+h}}{R^h} \leq W_t$, where R is the exogenous rate of return. The agent's objective expected utility is given by:

$$U_t(\vec{c_t}|\omega_t) = u(c_t|\omega_t) + \sum_{h=1}^{\infty} \beta^h \cdot \left[\sum_{\omega_{t+h} \in \Omega} \pi(\omega_{t+h}|\omega_t) u(c_{t+h}|\omega_{t+h}) \right], \tag{8}$$

where β is the discount rate, and the agent's marginal utility of consumption is potentially state-dependent. Equation 8 implies a standard Euler equation:

$$1 = \beta R \cdot E_t \left[\frac{u'(c_{t+1}|\omega_{t+1})}{u'(c_t|\omega_t)} \right] = \beta R \cdot E_t[M(\omega_{t+1})], \tag{9}$$

where $M(\omega_{t+1}) \equiv \frac{u'(c_{t+1}|\omega_{t+1})}{u'(c_t|\omega_t)}$ is the marginal utility of consumption in state ω_{t+1} relative to the current marginal utility of consumption.

Context and cues The intertemporal objective in equation 8 is highly complex to evaluate: a rational agent needs to integrate utility all possible future states over all time horizons. We assume for our agent, states instead come to mind in proportion to their similarity to her context C_t, as in equation 2. Further we assume that the cues in context C_t at time t include the agent's current and past experienced states, each with a weight w.

$$C_t = \{\omega_{t-h}, w(\omega_{t-h})\}_{h \geq 0}. \tag{10}$$

These assumptions imply that rather than optimizing her true objective function, given by equation 8, the agent instead chooses c_t to optimize:

$$U_t^s(\vec{c_t}|\omega_t, C_t) \propto S(\omega_t, C_t) \cdot u(c_t|\omega_t) + \sum_{h=1}^{\infty} \beta^h \left[\sum_{\omega_{t+h} \in \Omega} \pi(\omega_{t+h}|\omega_t) S(\omega_{t+h}, C_t) u(c_{t+h}|\omega_{t+h}) \right]. \tag{11}$$

Proposition 2 then follows:

Proposition 2. *Let* $\tilde{S}(\omega_{t+1}) \equiv \frac{S(\omega_{t+1}, C_t)}{S(\omega_t, C_t)}$ *be the similarity of* ω_{t+1} *to* C_t*, normalized by similarity*

of ω_t to C_t. Equation 11 implies the following modified Euler equation:

$$\beta R \cdot E_t[M(\omega_{t+1})] = \left(E_t[\tilde{S}(\omega_{t+1})]\right)^{-1} - \beta R \cdot \mathcal{D}^M(C_t), \tag{12}$$

where $\mathcal{D}^{M(\omega_{t+1})}(C_t) = \frac{Cov_t[M(\omega_{t+1}),S(\omega_{t+1},C_t)]}{E_t[S(\omega_{t+1},C_t)]}$ is the representativeness of the current context with respect to future marginal utility.

Note that in the absence of similarity distortions, $\mathcal{D}^{M(\omega_{t+1})} = 0$ and $\tilde{S}(\omega_{t+1}) = 1$, which recovers the normative benchmark (equation 9). Under similarity-based distortions, Proposition 2 implies that there is excessively low current consumption relative to the normative benchmark (Equation 9) if the current context is representative of high marginal-utility states in the future (high $\mathcal{D}^{M(\omega_{t+1})}(C_t)$). Conversely, there is excessively high current consumption if future states are less associated with C_t (low $E_t[\tilde{S}(\omega_{t+1})]$). We discuss each channel separately.

Experience and projection effects The term $\mathcal{D}^M(C_t)$ captures whether the current context is disproportionately associated with high or low marginal utility future states. One can further decompose this term into the current state ω_t and past experiences:

$$\mathcal{D}^M(C_t) = \frac{w(\omega_t)\mathcal{W}(\omega_t)}{w(\omega_t)\mathcal{W}(\omega_t) + \mathcal{W}_{past}} \mathcal{D}^{M(\omega_{t+1})}(\omega_t) + \frac{\mathcal{W}_{past}}{w(\omega_t)\mathcal{W}(\omega_t) + \mathcal{W}_{past}} \mathcal{D}^{M(\omega_{t+1})}(C_t^{past}), \tag{13}$$

where $C_t^{past} = \{\omega_{t-h}, w(\omega_{t-h})\}_{h\geq 1}$ and $\mathcal{W}_{past} = \sum_{h\geq 1} w(\omega_{t-h}) \cdot \mathcal{W}(\omega_{t-h})$. Equation 13 shows how cueing and similarity can provide a unifying framework for two well-established findings in behavioral economics. First, it captures experience effects: $\mathcal{D}^{M(\omega_{t+1})}(C_t^{past})$ increases when individuals have experienced past economic disasters. Associations with past disaster states reflect the lasting imprint of lived experiences on current decision-making (Malmendier, 2021). Second, it captures projection bias: $\mathcal{D}^{M(\omega_{t+1})}(\omega_t)$ increases during transient adverse states, illustrating how individuals implicitly over-extrapolate current conditions (Loewenstein et al., 2003).

Beyond offering a shared framework for these effects, Equation 13 suggests that the degree to which past experiences and present conditions shape consumption decisions depends on similarity. In particular, similarity assessments can help researchers better understand the scope of experience effects: the theory naturally implies that experience effects are stronger in contexts that resemble or are closely associated with past experiences. Measuring the associations between past experiences and the current context is key to understanding the *domain-specificity of experience effects* – whether differences in experiences translate to strong differences in expectations in a novel setting (e.g. would

Depression babies react more strongly to inflation?).

Moreover, equation 13 highlights that experience effects and projection bias can function as *substitutes*, because they are driven by separate cues (components of the local context) that compete for individual attention. For example, older individuals displayed greater resilience to pessimism during the COVID-19 pandemic – their experiences of overcoming past adversity outweighed their current experiences of the Covid crisis (Bordalo, Burro, et al., 2024a). Conversely, the history of repeated financial crises are replete with circumstances where the current narrative ("this time is different") crowds out past experiences (or wisdom) of repeated crises (Reinhart, 2009). Our framework not only clarifies the interaction between past experiences and present context but also suggests avenues for future research to quantify these effects using similarity assessments.

Endogenous cognitive discounting The second term of equation 12, $(E_t[\tilde{S}(\omega_{t+1})])^{-1}$, captures how easy it is for the individual to simulate consumption in the future. The harder it is for the agent to think of the future (lower $E_t[\tilde{S}(\omega_{t+1})]$), consumption today is higher relative to the normative benchmark (excessively high future marginal utility of consumption). Our model is consistent with a body of work on cognitive discounting – that time preferences may reflect difficulty in thinking about the future (Becker & Mulligan, 1997; Gabaix & Laibson, 2017; Rick & Loewenstein, 2008). Relative to the literature, our framework emphasizes that the degree of cognitive discounting (and implicit time preferences) itself is *context-dependent* and depends on the similarity between the current state and future states of the world. In particular, our framework makes the novel prediction that there is greater impatience in *unusual circumstances* that are more dissimilar to normal states of the world.

To illustrate this, consider the simple case in which the utility of consumption $u(c_t|\omega_t)$ does not depend on ω_t. Despite being completely normatively irrelevant, ω_t nevertheless distorts consumption by influencing the relative availability of future consumption. Assume that ω_t follows a stationary AR(1) process: $\omega_t = \rho\omega_{t-1} + \epsilon_t$, $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$, where we impose as before the similarity function $S(\omega, \omega') = \exp\left(-\frac{\kappa}{2}(\omega - \omega')^2\right)$. Equation 11 reduces to the following expression.

Proposition 3. *The agent effectively optimizes:* $U_t = u(c_t) + \sum_{h\geq 1} \delta_h u(c_{t+h})$*, where the implicit decision weights* δ_h *are given by*

$$\delta_h = \beta^h \cdot \sqrt{\frac{\tau}{\tau+\kappa}} \cdot \exp\left(-\frac{1}{2}\left(\frac{\tau\cdot\kappa}{\tau+\kappa}\cdot(1-\rho^h)^2\omega^2\right)\right),\ \tau = \frac{1-\rho^2}{\sigma^2(1-\rho^{2(h+1)})} \tag{14}$$

Figure 6 plots the endogenous time discount factors δ_h. Relative to the normative

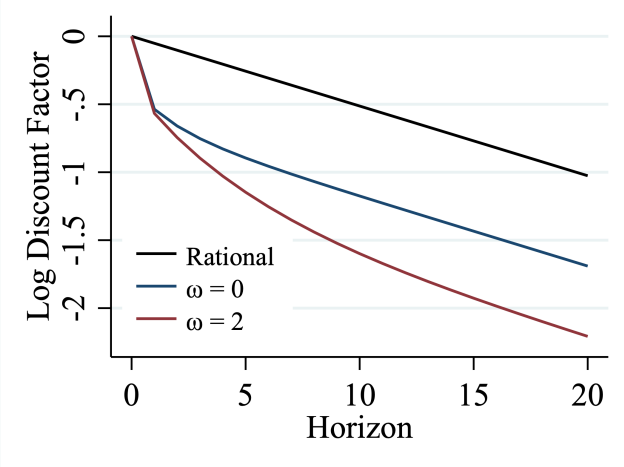


Figure 6: Endogenous cognitive discounting

benchmark (β^h) plotted in black, the discount factors have a sharper present bias. This arises endogenously due to the fact that the future states are less similar to current state. More importantly, the *degree* to which the future is discounted against the present is state dependent: there is much lower discounting when $\omega = 0$, the median state of the stationary distribution. Conversely, when $\omega = 2$ – the agent is experiencing a more "unusual" state – there is sharper discounting. For example, when individuals are going through a financial crisis, a health emergency, or emotional distress, they may over-prioritize the current exigent circumstances. Similarly, an agent may overspend while in a nice vacation (because the boring future is really not top of mind). Conversely, an agent may also be induced to prioritize future consumption, if it is in a setting that is similar to the current setting (e.g. selling the next vacation in the current vacation). All these examples highlight the context-dependence of time preferences.

Lastly, note that the log discount curves are convex: the relative discounting of consumption in $t+h$ to $t+h+1$ decreases as $h \mapsto \infty$. This "kink" in the time discounting function generates dynamic inconsistency in consumption behavior (Laibson, 1994). We endogenously obtain this from the simple insight that the far future and the very far future are intuitively equally (dis)similar to the present. Formally, for any stationary process ω_t, one obtains:

$$\lim_{h \mapsto \infty} E[S(C_t, \omega_{t+h})] \mapsto \sum_{\omega \in \Omega} \pi_{stat}(\omega) \cdot S(C_t, \omega), \tag{15}$$

where π_{stat} is the stationary distribution of ω_t. Thus, the relative discounting between $t+h$ and $t+h+1$ decreases to β, which generates endogenous hyperbolic discounting.

Summary To summarize, we have applied our framework to a simple intertemporal choice setting, where an agent's choice of consumption across time is distorted by both salient past experiences and the current state. In that case, similarity judgments across different states of the world help unite and shed new light on how local context distorts intertemporal decisionmaking.

4.2 Endogenous cues: application to persuasion

While the previous application has focused on *exogenous* cues – past and current experiences – we now turn to the case of *endogenous* cues, where a persuader chooses the optimal cue to distort receiver beliefs. We assume as before that the receiver is naive about the contextual distortion, and in particular does not take into consideration the strategic motives of the persuader, following other work in behavioral persuasion (Schwartzstein & Sunderam, 2021).

Setup Consider first the following general case. A persuader (who can be viewed as a monopolist firm) produces a single good for a receiver, which delivers $\psi(\omega)$ in state ω. We assume that $\psi(\omega)$ is known by the receiver: the persuader therefore cannot influence receiver through traditional persuasion, such as by providing objectively informative signals about the frequency of states or the quality of the good in them. Instead, it simply chooses a cue $Q \subset \Omega$ that alters the consumer's context when she considers purchasing the good. That is, it can run advertisements that highlight particular states, making them (and associated similar states) more likely to come to mind when its potential customers think about purchasing the good.[7]

As given in Proposition 1, the receiver's subjective willingness to pay is given by:

$$E_s[\psi|C \cup Q] = E[\psi] + \frac{\sum_{\omega_c \in C} w(\omega_c) \cdot \mathcal{W}(\omega_c) \cdot \mathcal{D}^{\psi}(\omega_c)}{\sum_{\omega_c \in C} w(\omega_c) \cdot \mathcal{W}(\omega_c)}. \tag{16}$$

We assume the persuader extracts the full surplus from the exchange: it is able to charge the receiver their willingness to pay for the good net of reserve utility ψ_0. The persuader thus solves a joint optimization problem over both the product design (ψ) and the optimal cue (Q), where it maximizes the receiver's subjective willingness to pay net of the cost of

[7] Of course, even uninformative advertisements in practice can do more than simply highlight states. For example, they might influence the choice set considered by the consumer (Abaluck & Adams-Prassl, 2021), as well as which features of the good are salient, which may change how the good is categorized (Mullainathan et al., 2008). We leave a more general treatment of the role of uninformative ads to future work.

producing the good $c(\psi)$:

$$\left(Q^*, \psi^*\right) = \underset{Q,\psi,E_s[\psi|C\cup Q]\geq\psi_0}{\operatorname{argmax}} \; E_s[\psi|C \cup Q] - c(\psi). \tag{17}$$

Associative cue and product choice Equation 17 implies that the persuader's choice of the underlying product design (ψ) tilts towards goods that are *complementary* with the technology of associative cues. For what kind of goods will there be an effective cue? By Cauchy-Schwarz inequality, the upper bound on the persuasiveness of a cue is given by the following expression:

$$E_s[\psi|C \cup Q] - E[\psi] \leq \mathcal{D}^{\psi}(C \cup Q) \leq \operatorname{Var}\left(\frac{S(\omega, C \cup Q)}{E[S(\omega, C \cup Q)]}\right)^{1/2} \cdot \operatorname{Var}\Big(\psi(\omega)\Big)^{1/2}. \tag{18}$$

The first term, $\operatorname{Var}\left(\frac{S(\omega,C\cup Q)}{E[S(\omega,C\cup Q)]}\right)^{1/2}$, depends on the associative structure of the state space Ω: there must be variation in how similar states are to each other for cues to ease simulation of some states more than others. In particular, as discussed in Section 2.3 and equation 6, associative cues favor goods that deliver high utility in *distinctive* states of the world. The second term, $\operatorname{Var}\Big(\psi(\omega)\Big)^{1/2}$, highlights that associative cues are naturally complementary with products that deliver high utility in specialized states of the world: variation in utility generates scope for an associative cue to overstate the value of the good. This force leads to more ex post regret by the consumer: she does not find herself in the high-utility states as often as it felt she would when she bought the good. Such an agent might find herself with a rarely-visited timeshare vacation home, stuck in rush-hour traffic in a high-speed sports car, or frequently paying for product warranties and travel insurance she predictably ends up failing to need.

Application: portfolio choice and underdiversification When applied to risky choice and portfolio selection, equation 18 highlights a novel tension between *risk aversion* and *persuasiveness*. On one hand, an asset with high variance in expected returns is risky and thus less desirable for the investor. On the other hand, for there to be scope for cue-based persuasion, there needs to be variation in returns. For example, an advisor can help a client to see the upside of a risky stock (and interfere with simulating its downsides), but no cue can boost the subjective benefit of a riskless asset beyond its fixed nominal return. When the latter force dominates, financial persuasion can induce pessimistic investors—ones who would not even participate in the stock market otherwise—to adopt risky portfolios that primarily payoff only in distinctive (easy to cue) states. But such an

advisor may not be able (and in any case will not choose) to convince her to pursue the rational diversified strategy. Below, we formalize these ideas by solving out the advisor's joint portfolio-and-cue choice in a simple two asset setting.

Let $\Omega = \{\omega_A, \omega_B, \omega_0\}$ with $\pi(\omega_A) = \pi(\omega_B) = p$. Each state of the world corresponds to the realization of a distinct investment "theme" – for example, there may be an upcoming AI revolution, or a breakthrough in battery technology, or neither ($\omega_0 =$ the status quo).[8] Correspondingly, there are two risky assets, also indexed by A and B, with the excess return on each asset given by:

$$R_A = \alpha \cdot I(\omega = \omega_A) + \epsilon_A,\ R_B = \alpha \cdot I(\omega = \omega_B) + \epsilon_B, \tag{19}$$

with ϵ_A and ϵ_B uncorrelated with variance σ^2. We endow Ω with the same similarity structure as that in Section 2.3: there are two features that govern the similarity between states: whether any investment theme has materialized or not, as well as which theme has materialized. This implies $S(\omega_0, \omega_A) = S(\omega_0, \omega_B) = \delta_0 \leq 1$, and $S(\omega_A, \omega_B) = \delta_1 \leq 1$.

Investors have mean-variance preferences over wealth $u(W) = E[W] - \frac{A}{2} Var[W]$, with access to a riskless saving technology with excess return $R_0 = 0$. They can access risky investments only through their financial advisor. This advisor first provides a cue $Q \subset \Omega$ that highlights some states of the world but does not provide objective information on their likelihood. In particular, the advisor has access to cues of total weight w_Q, which it can allocate to ω_A and ω_B: $Q^* = \{\omega_A, w^*_Q(\omega_A), \omega_B, w^*_Q(\omega_B)\}$. We assume for simplicity that the context (absent the persuader's cue) is in the "status quo" state: $C = \{\omega_0, w(\omega_0)\}$.[9] Along with the cue, the advisor jointly offers the investor a risky portfolio $x^* = (x^*_A, x^*_B)$, $x^*_A + x^*_B = 1$ at the cost of a fee ϕ^*, which the investor accepts if the subjective expected utility from them exceeds the reserve utility from the riskless asset.[10] Proposition 4 characterizes the optimal persuader action (Q^*, x^*):

Proposition 4. *The persuader adopts a pure cue* $\max\{w_{Q^*}(\omega_A), w_{Q^*}(\omega_B)\} = w_Q$ *and offers a portfolio* x^* *tilted towards the asset that loads on the cued state. The tilt of the portfolio* $T(x^*) \equiv \frac{\max(x^*_A, x^*_B)}{\min(x^*_A, x^*_B)}$ *is decreasing in the similarity* δ_1 *between the themes.*

[8] We assume for simplicity these themes cannot co-occur.

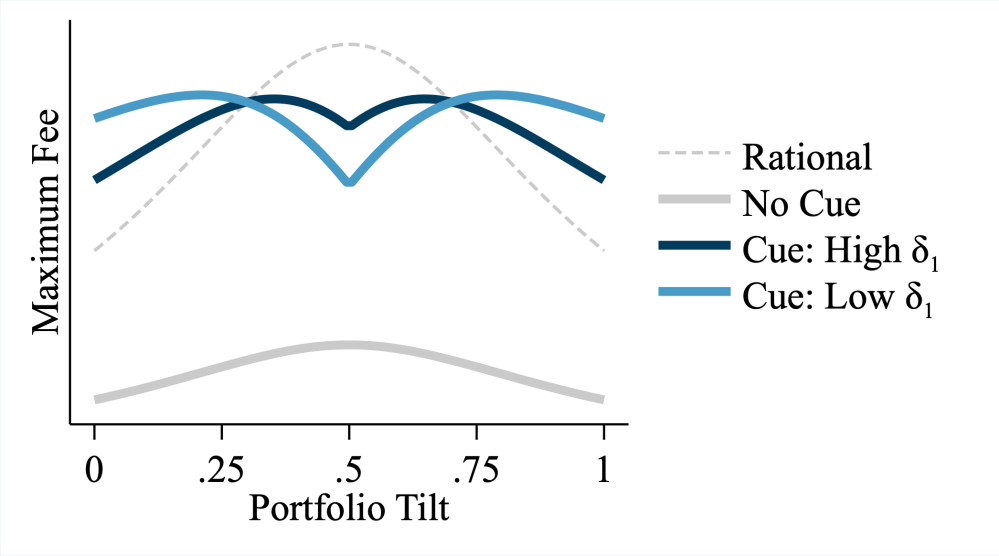
[9] Our persuasion results are unchanged if instead we set C to be the unbiased distribution over Ω, though of course what the investor would choose to do absent an advisor (e.g., whether she would choose not to participate) depends on the choice of default.

[10] Standard mean-variance optimization implies that the advisor offers a portfolio $x^* = \frac{1}{A} Var_s^{-1}[R](E_s[R] - R_0)$ and extracts fees

$$\phi^* = \frac{1}{A}(E_s[R] - R_0)' Var_s^{-1}[R_i](E_s[R] - R_0), \tag{20}$$

where the expectations and variance are taken over the investor's subjective distribution given the context: $\pi_s(\omega) \propto \pi(\omega) \cdot S(\omega, C \cup Q)$.

Figure 7: Financial persuasion: portfolio choice



To give an intuition for Proposition 4, Figure 7 plots the maximal fees obtained for a given portfolio (x_A, x_B) as one varies the tilt of the portfolio $(1,0)$ to $(0,1)$. For both the rational benchmark and the persuader-less benchmark, fees are maximized by the fully-diversified portfolio: investors value safety and reward advisors that deliver it. When the advisor can also provide a cue, however, it is the more extreme portfolios (that tilt towards either investment theme) that receive the most boost – as shown by the gap between the blue and gray curves. The intuition is exactly that of Equation 18: there exists a more effective cue for the less diversified portfolio. The advisor's optimal strategy thus tilts the portfolio towards a particular investment theme, optimally trading off persuasiveness with losses from lack of diversification. Furthermore, the degree of diversification depends on the associativity between different investment opportunities: the more similar the themes, the more willing investors are to invest in both opportunities. In particular, under the natural assumption that associativity is positively related to objective correlation—assets that do well in similar circumstances are perceived as more similar—investors will be more persuadable to jointly invest in highly correlated projects, the opposite of what rational diversification would suggest. More generally, this suggests that measuring the associativity between different investment outcomes may be important to understanding financial persuasion and the resulting investment behavior.

Comparisons with informative persuasion Lastly, it is useful to compare our model to a benchmark of purely informative persuasion. Consider the case where the persuader

can provide true information at a cost, with the costs proportional to the mutual information between the prior and the posterior (Sims 2011). Formally, the informative persuader maximizes over local perturbations $(\pi(\omega_A), \pi(\omega_B)) = (\pi_0(\omega_A) + d\pi(\omega_A), \pi_0(\omega_B)) + d\pi(\omega_B))$ as follows:

$$\max \phi^*(\pi(\omega_A), \pi(\omega_B)) - \xi \left[\log\left(\frac{\pi(\omega_A)}{\pi_0(\omega_A)}\right)\pi(\omega_A) + \left(\frac{\pi(\omega_B)}{\pi_0(\omega_B)}\right)\pi(\omega_B) + \left(\frac{\pi(\omega_0)}{\pi_0(\omega_0)}\right)\pi(\omega_0) \right], \quad (21)$$

where $\phi^*(\pi(\omega_A), \pi(\omega_B))$ are optimal fees given investor beliefs. In that case, the persuader instead chooses the following strategy:

Proposition 5. *The informative persuader locally boosts π_A and π_B equally and offers a perfectly diversified portfolio: $T(x^*) = 1$.*

Intuitively, the persuader would like to pay the information costs and increase the expected returns of both themes. This is because the informative persuader faces different constraints from one relying only on cues. While the informative persuader can boost the subjective returns of both assets, an advisor that relies just on cues faces the constraint that different cues distract from each other. In that case, simple (and risky) is best: better to pick a theme and maximally boost its simulation.

5 Extensions: information and cues

Finally, while we have primarily focused on the case of purely uninformative cues for conceptual clarity, we explore how our framework generalizes to settings where there is also genuine information regarding Ω. The generalized framework yields novel implications on how informative signals influence the effectiveness of cues.

Setup Suppose that the agent also learns of an objective signal I associated with the likelihood function $\pi(I|\omega) : \Omega \mapsto \mathbb{R}$. We assume that the agent reacts to I also by sampling states $\omega \in \Omega$ subject to the local context C. That is, given both the context C and information I, the DM's subjective distribution across Ω is given by:

$$\pi(\omega|C, I) \propto S(\omega, C) \cdot \pi(\omega) \cdot \pi(I|\omega). \quad (22)$$

To give an intuition behind Equation 22, just like how the DM was previously assessing a known distribution $\pi(\omega)$ subject to sampling distortions, the DM is now reacting to the signal in a Bayesian manner – combining the prior and the likelihood – subject to the

same cognitive constraint as before.[11] Equation 22 reduces to our main framework when I is completely uninformative ($\pi(I|\omega) = 1$ for all ω). Conversely, it also nests Bayesian updating when one removes similarity distortions ($S(\omega, C) = 1$).

Persuasion and information To further explore the implication of equation 22, consider the following simple example. Let $\Omega = \mathbb{R}$, with $S(\omega, \omega') = \exp(-\frac{\kappa}{2}(\omega - \omega')^2)$, with the objective distribution given by $\pi(\omega) = \phi_{\mathcal{N}}(\mu_0, \tau_0^{-1})$. Suppose the persuader wants to maximize $E_s[\psi(\omega)] \equiv E_s[\omega]$, and has two means at her disposal. First, as before, the persuader can give an uninformative cue Q. Second, the persuader now has access to a noisy signal I of ω with precision τ, with the likelihood function $\pi(I|\omega) = \phi_{\mathcal{N}}(\omega, \tau^{-1})$. In that case, the Bayesian posterior is given by:

$$\pi(\omega|I) = \phi_{\mathcal{N}}\left(\frac{\tau_0\mu_0 + \tau I}{\tau_0 + \tau}, (\tau_0 + \tau)^{-1}\right) \equiv \phi_{\mathcal{N}}\left(\mu_{post}, \tau_{post}^{-1}\right) \tag{23}$$

To complete the setup, we assume for simplicity that $\mathcal{D}_d = \mu_{post}$: this implies that the subjective mean in absence of the persuader is equal to the objective posterior – this is to ensure that any distortion in beliefs is purely due to the persuader-provided cue. Given these assumptions, one obtains the following closed-form expression for the optimal persuader cue Q^*:

Proposition 6 (Optimal cue)**.** *The persuader's optimal cue Q^* and agent beliefs are given by:*

$$Q^* = \mu_{post} + \sqrt{\frac{\tau_{post} + \kappa}{\tau_{post} \cdot \kappa}\nu_0} \quad \text{and} \quad E[\omega|Q^*, I] = \mu_{post} + \sqrt{\frac{\kappa}{\tau_{post} \cdot (\tau_{post} + \kappa)}}\left(\sqrt{\nu_0} - \frac{1}{\sqrt{\nu_0}}\right), \tag{24}$$

where $\nu_0 > 1$ is the unique solution to $1 + \frac{1}{W_d}\exp\left(-\frac{1}{2}\nu_0\right) = \nu_0$.

Corollary 1. *Agent beliefs $E[\omega|Q^*, I]$ are decreasing in τ_{post} and increasing in κ. In particular, for any signal I with precision τ, there exists an alternative signal $I' > I$ with lower precision $\tau' < \tau$ that the persuader prefers, despite it resulting in a lower objective posterior mean.*

Corollary 1 highlights the potential tradeoff faced by a persuader who has access to both information and cues. On the one hand, a positive signal I of high precision will move objective beliefs strongly in the persuader's favored direction. On the other hand, as highlighted by Equation 24, a signal of higher precision limits the scope for persuasion through cues: the more precise the beliefs, the lower the availability of cues that stray

[11] For parsimony, we are not modeling the possibility that the information, or the data, itself can act as a cue, which causes the receiver to sample near information consistent with the data. We explore this possibility in previous work (Bordalo et al., 2023b).

from the objective posterior. Thus, rather than providing a positive signal with high precision, the persuader may instead find it more effective to provide a more extreme signal with low precision, while simultaneously cuing extreme states consistent with the signal. Our results thus suggest another mechanism for why there may be greater overreaction to weak or extreme signals (Kwon & Tang, 2020; Afrouzi et al., 2023; Bohren et al., 2024; Augenblick et al., 2024): these are signals for which a persuader can find a particularly effective cue.

6 Conclusion

This paper builds on two central premises. First, the set of relevant states for a decision is often vast, and agents may therefore be unable to consider and integrate over all of them before choosing. They must rely instead on the states they manage to simulate. The second premise is that this process of simulation is context-dependent and therefore susceptible to cues. Our experiments corroborate these ideas and highlight the role that similarity and interference play in shaping how irrelevant contextual cues distort beliefs. They also point toward a novel (to economics) type of data—similarity judgments—as a method of generating predictions and disciplining behavioral models.

Two directions for future work appear promising. First, in this paper we focus on an exogenously given and fixed associative structure between states. This is obviously an oversimplification, as the perceived similarity between objects is a function of the features agents happen to be attending to (Tversky 1977). Introducing endogenous attention, and therefore similarity, would enrich the model: cues might not only shape which states initially come to mind, but also which features agents attend to while evaluating their options, which in turn may shape the associations between different states (see Bordalo, Gennaioli, et al. 2024a for related ideas). Second, the usefulness of similarity data in predicting treatment effects in the lab raises the possibility of measuring associations in the field. Which states come to mind when considering a financial investment, or which situations one imagines using a product in, depend on those states' similarity structures. Measuring and deploying similarity judgments to map out these associations and the corresponding effects that cues have "in the wild" are open and exciting avenues for future research.

References

Abaluck, J., & Adams-Prassl, A. (2021). What do consumers consider before they choose? identification from asymmetric demand responses. *The Quarterly Journal of Economics*, *136*(3), 1611–1663.

Afrouzi, H., Kwon, S. Y., Landier, A., Ma, Y., & Thesmar, D. (2023). Overreaction in expectations: Evidence and theory. *The Quarterly Journal of Economics*, *138*(3), 1713–1764.

Alesina, A., & Giuliano, P. (2011). Preferences for redistribution. In *Handbook of social economics* (Vol. 1, pp. 93–131). Elsevier.

Alesina, A., Miano, A., & Stantcheva, S. (2023). Immigration and redistribution. *The Review of Economic Studies*, *90*(1), 1–39.

Ashby, F. G., & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review*, *95*, 124-150. doi: 10.1037/0033-295X.95.1.124

Augenblick, N., Backus, M., Little, A., & Moore, D. (2024). *Model uncertainty, disagreement, and over-precision: Theory and evidence.* (Working paper)

Ba, C., Bohren, J. A., & Imas, A. (2023). *Over-and underreaction to information.* Retrieved from https://ssrn.com/abstract=4274617

Barberis, N. C., & Jin, L. J. (2023). *Model-free and model-based learning as joint drivers of investor behavior* (Tech. Rep.). National Bureau of Economic Research.

Barron, K., & Fries, T. (2023). Narrative persuasion. *CESifo Working Paper no. 10206*.

Becker, G. S., & Mulligan, C. B. (1997). The endogenous determination of time preference. *The Quarterly Journal of Economics*. Retrieved from https://academic.oup.com/qje/article/112/3/729/1926899

Becker, G. S., & Murphy, K. M. (1993). A simple theory of advertising as a good or bad. *The Quarterly Journal of Economics*, *108*(4), 941–964.

Benoit, R. G., & Schacter, D. L. (2015). Specifying the core network supporting episodic simulation and episodic memory by activation likelihood estimation. *Neuropsychologia*, *75*, 450-457. Retrieved from http://dx.doi.org/10.1016/j.neuropsychologia.2015.06.034 doi: 10.1016/j.neuropsychologia.2015.06.034

Bohren, J. A., Hascher, J., Imas, A., Ungeheuer, M., & Weber, M. (2024). *A cognitive foundation for perceiving uncertainty.* Retrieved from https://ssrn.com/abstract=4706147

Bordalo, P., Burro, G., Coffman, K., Gennaioli, N., & Shleifer, A. (2024a). Imagining the future: memory, simulation, and beliefs. *Review of Economic Studies*, rdae070.

Bordalo, P., Burro, G., Coffman, K., Gennaioli, N., & Shleifer, A. (2024b). Imagining the future: Memory, simulation, and beliefs about covid. *Review of Economic Studies*.

Bordalo, P., Conlon, J. J., Gennaioli, N., Kwon, S. Y., & Shleifer, A. (2023a). How people use statistics. *NBER Working Paper No. 31631*.

Bordalo, P., Conlon, J. J., Gennaioli, N., Kwon, S. Y., & Shleifer, A. (2023b). Memory and probability. *The Quarterly Journal of Economics*, *138*(1), 265–311.

Bordalo, P., Gennaioli, N., Lanzani, G., & Shleifer, A. (2024a). A cognitive theory of reasoning and choice. *Working paper*.

Bordalo, P., Gennaioli, N., Lanzani, G., & Shleifer, A. (2024b). A cognitive theory of reasoning and choice.

Caplin, A., Dean, M., & Leahy, J. (2019). Rational inattention, optimal consideration sets, and stochastic choice. *The Review of Economic Studies*, *86*(3), 1061–1094.

Charles, C. (2022a). *Memory and trading.*

Charles, C. (2022b). *Memory moves markets.*

Charles, C., & Kendall, C. (2024). Causal narratives. *Working paper*.

Chater, N., Zhu, J. Q., Spicer, J., Sundh, J., León-Villagrá, P., & Sanborn, A. (2020, 10). Probabilistic biases meet the bayesian brain. *Current Directions in Psychological Science*, *29*, 506-512. doi: 10.1177/0963721420954801

Chivers, T. (2019, 12). A theory in crisis. *Nature*.

Cohn, A., & Maréchal, M. A. (2016, 12). *Priming in economics* (Vol. 12). Elsevier B.V. doi: 10.1016/j.copsyc.2016.04.019

Colonnelli, E., Gormsen, N. J., & Mcquade, T. (2024). Selfish corporations. *The Review of Economic Studies*.

Conlon, J. (2024). Attention, Information, and Persausion. *Working paper*.

Dougherty, M. R., Gettys, C. F., & Thomas, R. P. (1997). The role of mental simulation in judgments of likelihood. *Organizational Behavior and Human Decision Processes*, *70*, 135-148. doi: 10.1006/obhd.1997.2700

Eliaz, K., & Spiegler, R. (2020). A model of competing narratives. *American Economic Review*, *110*, 3786-3816. doi: 10.2307/26966480

Enke, B., & Graeber, T. (2023). Cognitive uncertainty. *The Quarterly Journal of Economics*.

Enke, B., Schwerter, F., & Zimmerman, F. (2021). *Associative memory and belief formation.*

Enke, B., Schwerter, F., & Zimmermann, F. (2024). Associative memory, beliefs and market interactions. *Journal of Financial Economics*, *157*, 103853.

Esponda, I., & Vespa, E. (2014). Hypothetical thinking and information extraction in the laboratory. *American Economic Journal: Microeconomics*, *6*, 180-202. doi: 10.1257/mic.6.4.180

Evers, E. R. K., Imas, A., & Kang, C. (2021). On the role of similarity in mental accounting and hedonic editing. *Psychological Review*.

Fudenberg, D., Kleinberg, J., Liang, A., & Mullainathan, S. (2019). Measuring the completeness of theories. *arXiv preprint arXiv:1910.07022*.

Gabaix, X. (2019). Behavioral inattention. In *Handbook of behavioral economics: Applications and foundations 1* (Vol. 2, pp. 261–343). Elsevier.

Gabaix, X., & Laibson, D. (2017). *Myopia and discounting* (Tech. Rep.). National bureau of economic research.

Gabaix, X., & Laibson, D. (2022). Myopia and discounting. *NBER Working Paper 23254*. Retrieved from http://www.nber.org/papers/w23254

Gershman, S. J., Zhou, J., & Kommers, C. (2017, 12). Imaginative reinforcement learning: Computational principles and neural mechanisms. *Journal of Cognitive Neuroscience*, *29*, 2103-2113. doi: 10.1162/jocn_a_01170

Gilbert, D. T., & Wilson, T. D. (2007, 9). Prospection: Experiencing the future. *Science*, *317*, 1351-1354. doi: 10.1126/science.1140734

Gilboa, I., & Schmeidler, D. (1995). Case-based decision theory. *The Quarterly Journal of Economics*. Retrieved from https://academic.oup.com/qje/article/110/3/605/1859208

Graeber, T., Roth, C., & Zimmermann, F. (2023). *Stories, statistics, and memory*.

Haaland, I., Roth, C., & Wohlfart, J. (2023, 3). Designing information provision experiments. *Journal of Economic Literature*, *61*, 3-40. doi: 10.1257/jel.20211658

Hvidberg, K. B., Kreiner, C. T., & Stantcheva, S. (2023). Social positions and fairness views on inequality. *Review of Economic Studies*, *90*(6), 3083–3118.

Kahana, M. J. (2012). *Foundations of human memory*. OUP USA.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology*, *3*(3), 430–454.

Kahneman, D., & Tversky, A. (1981). The simulation heuristic. *Working paper.1*.

Kwon, S. Y., & Tang, J. (2020). Extreme events and overreaction to news. *Available at SSRN 3724420*.

Laibson, D. (1994). *Hyperbolic discounting and consumption* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics*, *112*(2), 443–478.

Link, S., Peichl, A., Roth, C., & Wohlfart, J. (2023). Attention to the macroeconomy. *ECONtribute Discussion Paper No. 256*. Retrieved from www.econtribute.de

Loewenstein, G., O'Donoghue, T., & Rabin, M. (2003). Projection bias in predicting future utility. *the Quarterly Journal of economics*, 1209–1248.

Malmendier, U. (2021). Exposure, Experience, and Expertise: Why Personal Histories Matter in Economics. *Journal of the European Economic Association*.

Malmendier, U., & Nagel, S. (2011). Depression babies: do macroeconomic experiences affect risk taking? *The quarterly journal of economics*, *126*(1), 373–416.

Martínez-Marquina, A., Niederle, M., & Vespa, E. (2019). Failures in contingent reasoning. *American Economic Review*, *109*, 3437-3474. doi: 10.2307/26789063

Milgrom, P., & Roberts, J. (1986). Price and advertising signals of product quality. *Journal of Political Economy*, *94*, 796-821.

Moser, J. (2019, 7). Hypothetical thinking and the winner's curse: an experimental investigation. *Theory and Decision*, *87*, 17-56. doi: 10.1007/s11238-019-09693-9

Mullainathan, S. (2002). A memory-based model of bounded rationality. *The Quarterly Journal of Economics*. Retrieved from https://academic.oup.com/qje/article/117/3/735/1932979

Mullainathan, S., Schwartzstein, J., & Shleifer, A. (2008). Coarse thinking and persuasion. *The Quarterly Journal Of Economics*.

Mullally, S. L., & Maguire, E. A. (2014). *Memory, imagination, and predicting the future: A common brain mechanism?* (Vol. 20). SAGE Publications Inc. doi: 10.1177/1073858413495091

Möbius, M. M., Niederle, M., Niehaus, P., & Rosenblat, T. S. (2022, 11). Managing self-confidence: Theory and experimental evidence. *Management Science*, *68*, 7793-77817. doi: 10.1287/mnsc.2021.4294

Niederle, M., & Vespa, E. (2023). Cognitive limitations: Failures of contingent thinking. *Annual Review of Economics*, *15*, 307-328. Retrieved from https://doi.org/10.1146/annurev-economics- doi: 10.1146/annurev-economics

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: learning, memory, and cognition*, *14*(1), 54.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annu. Rev. Psychol*, *43*, 25-53. Retrieved from www.annualreviews.org

O'Donoghue, T., & Rabin, M. (1999). Doing it now or later. *American economic review*, *89*(1), 103–124.

Oprea, R. (2024). Decisions under risk are decisions under complexity. *Working paper*.

Reinhart, C. M. (2009). *This time is different: Eight centuries of financial folly*. Princeton University Press.

Rick, S., & Loewenstein, G. (2008). Intangibility in intertemporal choice. *Philosophical transactions of the royal Society B: Biological Sciences*, *363*(1511), 3813–3824.

Sarkar, S. K. (2024). Economic representations.

Schacter, D. L., Addis, D. R., & Buckner, R. L. (2007, 9). *Remembering the past to imagine the future: The prospective brain* (Vol. 8). doi: 10.1038/nrn2213

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012, 11). *The future of memory: Remembering, imagining, and the brain* (Vol. 76). doi: 10.1016/j.neuron.2012.11.001

Schwartzstein, J., & Sunderam, A. (2021). Using models to persuade. *American Economic Review*, *111*, 276-323. doi: 10.1257/aer.20191074

Sims, C. A. (2011). Rational inattention and monetary economics. *Handbook of Monetary Economics*, *3A*. doi: 10.1016/S0169-7218(11)03004-8

Stantcheva, S. (2022). *How to run surveys: A guide to creating your own identifying variation and revealing the invisible *.* Retrieved from https://amerispeak.norc.org/us/en/amerispeak/about-amerispeak/overview.html

Stern, B. L., & Resnik, A. J. (1991). Information content in television advertising: A replication and extension. *Journal of Advertising Research*, *31*(3), 36–46.

Taubinsky, D., Butera, L., Saccarola, M., & Lian, C. (2024). *Beliefs about the economy are excessively sensitive to household-level shocks: Evidence from linked survey and administrative data *.*

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review*, *89*, 123-154.

Woodford, M. (2020). Modeling imprecision in perception, valuation, and choice. *Annual Review of Economics*, *12*, 579–601. doi: 10.1146/annurev-economics-102819-040518

A Proofs

Proof of Proposition 1 The expression follows directly from summing the simulation strength of each $\omega \in \Omega$:

$$\begin{aligned}\pi(H) &= \frac{\sum_{\omega\in H}\pi(\omega)}{\sum_{\omega\in\Omega}\pi(\omega)} = \frac{\sum_{\omega\in H} S(\omega,d)+\gamma S(\omega,Q)}{\sum_{\omega\in\Omega} S(\omega,d)+\gamma S(\omega,Q)} \\ &= \frac{1}{A(d)+\gamma A(Q)}\cdot\left(A(d)\cdot D(d)+\gamma A(Q)\cdot D(Q)\right)\end{aligned} \tag{25}$$

Proof of Proposition 2 To prove this proposition, one needs the following lemmas regarding cues on the real line. Recall that we assume the following similarity structure: $S(\omega,\omega') = \exp\left(-\frac{\kappa}{2}(\omega-\omega')^2\right)$. Note that given any normal density $\phi_{\mathcal{N}}(\mu_0,\tau_0^{-1})$, the density distorted by the cue Q is proportional to (excluding all terms unrelated to μ_0 and Q):

$$\exp\left[-\frac{\tau_0}{2}(\omega-\mu_0)^2-\frac{\kappa}{2}(\omega-Q)^2\right] = \exp\left[-\frac{\tau_0+\kappa}{2}\left(\omega-\frac{\tau_0\mu_0+\kappa Q}{\tau_0+\kappa}\right)^2-\frac{1}{2}\frac{\kappa\tau_0(\mu_0-Q)^2}{\tau_0+\kappa}\right]. \tag{26}$$

This implies that for a cue (d,Q), the subjective distribution is a mixture of two normal distributions $\phi_{\mathcal{N}}\left(\frac{\tau_0\mu_0+\kappa Q}{\tau_0+\kappa},(\tau_0+\kappa)^{-1}\right)$, $\phi_{\mathcal{N}}\left(\frac{\tau_0\mu_0+\kappa d}{\tau_0+\kappa},(\tau_0+\kappa)^{-1}\right)$ each with weight proportional to $\gamma\cdot\exp\left(-\frac{1}{2}\frac{\kappa\tau_0(\mu_0-Q)^2}{\tau_0+\kappa}\right)$ and $\exp\left(-\frac{1}{2}\frac{\kappa\tau_0(\mu_0-d)^2}{\tau_0+\kappa}\right)$. As assumed in the proposition, we focus on the case where $d=\mu_0$: the default context does not on its own introduce belief distortion. Then, the subjective beliefs simplify to:

$$E_s[\omega] = \mu_0 + \frac{\gamma\exp\left(-\frac{1}{2}\frac{\kappa\tau_0(\mu_0-Q)^2}{\tau_0+\kappa}\right)}{1+\gamma\exp\left(-\frac{1}{2}\frac{\kappa\tau_0(\mu_0-Q)^2}{\tau_0+\kappa}\right)}\cdot\frac{\kappa}{\tau_0+\kappa}(Q-\mu_0) \tag{27}$$

Setting $\eta_0 \equiv \frac{\kappa\tau_0}{\tau_0+\kappa}$ and $x \equiv Q-\mu_0$, it suffices to maximize (in x) the expression:

$$-\frac{1}{2}\eta_0 x^2 - \log\left(1+\gamma\exp\left(-\frac{1}{2}\eta_0 x^2\right)\right)+\log(x), \tag{28}$$

which generates the FOC:

$$0 = -\eta_0 x + \frac{1}{x} + \frac{\gamma\exp\left(-\frac{1}{2}\eta_0 x^2\right)\cdot\eta_0 x}{1+\gamma\exp\left(-\frac{1}{2}\eta_0 x^2\right)} \implies 1 = \frac{1}{1+\gamma\exp\left(-\frac{1}{2}\eta_0 x^2\right)}\eta_0 x^2, \tag{29}$$

which is satisfied when $\eta_0 x^2 = \nu_0$. Thus, $x = \sqrt{\nu_0/\eta_0} = \sqrt{\frac{\tau_{post}+\kappa}{\tau_{post}\kappa}\nu_0}k$, as desired. The expression for $E[\omega|Q^*, I]$ is immediate from substituting this expression into Equation 27.

Proof of Corollary 1 This is immediate where $\sqrt{\frac{\kappa}{\tau_{post}\cdot(\tau_{post}+\kappa}}$ decreasing in τ_{post} and increasing in κ.

Proof of Proposition 3 The expression for beliefs follow analogously from the derivation of Proposition 1. To prove the comparative static, note that for $A, B, A+A_0, B+B_0 > 0$, $\frac{A+A_0}{B+B_0} > \frac{A}{B} \iff A_0 > B_0\frac{A}{B}$. Applying this to the expression for beliefs, $\pi_{n+1} > \pi_n$ if and only if:

$$\pi_n \cdot \left(\tilde{\gamma}_{n+1}A(q_{n+1}) + (\tilde{\gamma}_{n+1} - \tilde{\gamma}_n)\sum_{k=1}^{n} A(q_k)\right) < \tilde{\gamma}_{n+1}A(q_{n+1})D(q_{n+1}) + (\tilde{\gamma}_{n+1} - \tilde{\gamma}_n)\sum_{k=1}^{n} A(q_k)D(q_k) \tag{30}$$

Substituting in: $\tilde{\gamma}_n \cdot \sum_{k=1}^{n} A(q_k)D(q_k) = \pi_n(A(d) + \tilde{\gamma}_n \sum_{k=1}^{n} A(q_k)) - A(d)\cdot\pi_d$, suffices to show:

$$\begin{aligned} &\pi_n \cdot \left(\tilde{\gamma}_{n+1}A(q_{n+1}) + (\tilde{\gamma}_{n+1} - \tilde{\gamma}_n)\sum_{k=1}^{n} A(q_k)\right) \\ &< \tilde{\gamma}_{n+1}A(q_{n+1})D(q_{n+1}) + \frac{\tilde{\gamma}_{n+1} - \tilde{\gamma}_n}{\tilde{\gamma}_n}\left(\pi_n(A(d) + \tilde{\gamma}_n \sum_{k=1}^{n} A(q_k)) - A(d)\cdot\pi_d\right). \end{aligned} \tag{31}$$

Rearranging, the above is equivalent to:

$$\begin{aligned} 0 &< \tilde{\gamma}_{n+1}A(q_{n+1})(D(q_{n+1}) - \pi_n) + \frac{\tilde{\gamma}_{n+1} - \tilde{\gamma}_n}{\tilde{\gamma}_n}(A(d)\cdot(\pi_n - \pi_d)). \\ &\iff \frac{\tilde{\gamma}_n - \tilde{\gamma}_{n+1}}{\tilde{\gamma}_n\tilde{\gamma}_{n+1}}(A(d)\cdot(\pi_n - \pi_d)) < A(q_{n+1})(D(q_{n+1}) - \pi_n), \end{aligned} \tag{32}$$

as desired.

Proof of Corollary 2 The proof is immediate from the fact that beliefs are convex combinations of the diagnosticity of the default context and each of the cues, with the weight on the default context converging to $1 - w^* = \frac{1}{1+\gamma(n)}$.

B Application proofs

B.1 Financial persuasion

We require the following lemmas:

Lemma 2. *Denote $\pi_A \equiv \sum_{\omega \in H_A} \pi(\omega|d,Q)$ and $\pi_B \equiv \sum_{\omega \in H_B} \pi(\omega|d,Q)$. The set of beliefs that the persuader can induce is given by (π_A, π_B) such that:*

$$\pi_A + \pi_B = \frac{2\delta_0 + \gamma(1+\delta_1)}{1 + 2\delta_0 + \gamma(1+\delta_0+\delta_1)} p, \ \min\{\pi_A, \pi_B\} \geq \frac{\delta_0 + \delta_1 \gamma}{1 + 2\delta_0 + \gamma(1+\delta_0+\delta_1)} p \tag{33}$$

Proof. This follows from the fact that the beliefs induced by the cue (γ_1, γ_2) is given by:

$$(\pi_1, \pi_2, \pi_0) \propto (\gamma_1 + \gamma_2\delta_1 + \delta_0, \gamma_2 + \gamma_1\delta_1 + \delta_0, 1 + \bar{\gamma}\delta_0). \tag{34}$$

□

Lemma 3. *Suppose that the persuader is constrained to offer a portfolio with fixed weight $(x_1, x_2) = W \cdot (\eta_1, \eta_2)$, where $\eta_1 + \eta_2 = 1$ and $\eta_i \geq 0$. The persuader's message consists of $(\gamma_1, \gamma_2)^* = (\bar{\gamma}, 0)$ if $\eta_1 > \eta_2$ and $(0, \bar{\gamma})$ otherwise.*

Proof. Assume without loss of generality that $\eta_1 \geq \eta_2$. First, we solve for the mean and variance of the returns given beliefs (π_0, π_1, π_2). We apply the law of iterated covariance:

$$Cov(X,Y) = E[Cov(X,Y|Z)] + Cov[E[X|Z], E[Y|Z]], \tag{35}$$

where X and Y are R_i, and Z is the realization of the stochastic state ω. One can work out that given the above set-up, the perceived covariance of returns for $[R_1, R_2]$ is given by:

$$E[R] = \alpha \cdot \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix}, \ Var[R] = \begin{bmatrix} \sigma^2 + \alpha^2\pi_1(1-\pi_1) & -\alpha^2\pi_1\pi_2 \\ -\alpha^2\pi_1\pi_2 & \sigma^2 + \alpha^2\pi_2(1-\pi_2), \end{bmatrix} \tag{36}$$

Consider a portfolio with weight η_i in asset i with $\eta_1 + \eta_2 = 1$. The maximal fee for

these relative weights is given by:

$$f^*(\eta_1,\eta_2,\pi_1,\pi_2) = \max_W W(\eta_1\pi_1+\eta_2\pi_2)\alpha - \frac{A}{2}W^2\begin{bmatrix}\eta_1 & \eta_2\end{bmatrix}Var[R]\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix} = \frac{1}{2A}\frac{(\eta_1\pi_1+\eta_2\pi_2)^2\alpha^2}{\begin{bmatrix}\eta_1 & \eta_2\end{bmatrix}Var[R]\begin{bmatrix}\eta_1\\ \eta_2\end{bmatrix}}$$
$$= \frac{1}{2A}\frac{(\eta_1\pi_1+\eta_2\pi_2)^2\alpha^2}{(\sigma^2+\alpha^2\pi_1(1-\pi_1))\eta_1^2+(\sigma^2+\alpha^2\pi_2(1-\pi_2))\eta_2^2-2\alpha^2\pi_1\pi_2\eta_1\eta_2}$$
$$= \frac{1}{2A}\frac{(\eta_1\pi_1+\eta_2\pi_2)^2\alpha^2}{\sigma^2(\eta_1^2+\eta_2^2)+(\pi_1\eta_1^2+\pi_2\eta_2^2)\alpha^2-(\eta_1\pi_1+\eta_2\pi_2)^2\alpha^2} \tag{37}$$

Given that $f(x) = 1/(x^{-1}+1)$ is monotonically increasing in x, $f^*(\eta_1,\eta_2)$ is maximized if and only if:

$$\frac{(\eta_1\pi_1+\eta_2\pi_2)^2\alpha^2}{\sigma^2(\eta_1^2+\eta_2^2)+(\pi_1\eta_1^2+\pi_2\eta_2^2)\alpha^2} \tag{38}$$

is maximized.

By Lemma 2, suffices to consider the case in which $\pi_1+\pi_2=\bar{\pi}$, with π_1,π_2 ranging over $\bar{\pi}\cdot(1-\pi^*,\pi^*)$ where $\pi^*>\frac{1}{2}$. We proceed with the proof in two steps. First, we prove that the expression in Equation 38 is U-shaped: increasing for $\pi_1 > C$ and decreasing for $\pi_1 < C$. Second, we then compare expression for the lowest and highest possible values of π_1: $\pi_1 = \bar{\pi}\pi^*$ and $\bar{\pi}(1-\pi^*)$ to show that the former is higher. Put together, this implies that the expression is globally maximized at π_1 at its maximal value (which corresponds to $(\gamma_1,\gamma_2) = (\bar{\gamma},0)$.).

Substituting in $\pi_2 = \bar{\pi}-\pi_1$, and taking the derivative with respect to π_1, we obtain that the expression is positive if and only if:

$$2(\eta_1-\eta_2)((\eta_1-\eta_2)\pi_1+\eta_2\bar{\pi})\cdot(\kappa(\eta_1^2+\eta_2^2)+(\bar{\pi}\eta_2^2+(\eta_1^2-\eta_2^2)\pi_1))-((\eta_1-\eta_2)\pi_1+\eta_2\bar{\pi})^2\cdot(\eta_1^2-\eta_2^2)>0$$
$$\iff 2\kappa(\eta_1-\eta_2)(\eta_1^2+\eta_2^2)+2(\eta_1-\eta_2)(\eta_1^2\pi_1+\eta_2^2(\bar{\pi}-\pi_1))-(\eta_1\pi_1+\eta_2(\bar{\pi}-\pi_1))(\eta_1^2-\eta_2^2)>0 \tag{39}$$

The expression is linear in π_1 with the slope of π_1 given by $(\eta_1-\eta_2)(\eta_1^2-\eta_2^2)>0$: thus, the expression is U-shaped (which includes monotonically increasing or decreasing). Thus, to find the global maximum, suffices to compare the original function evaluated at the end points: $\pi_1=\pi_1^*>\frac{1}{2}$ and $\pi_1 = 1-\pi_1^* = \pi_2^*$. Thus, suffices to show:

$$\frac{(\eta_1\pi_1^* + \eta_2\pi_2^*)^2}{\kappa(\eta_1^2+\eta_2^2)+(\eta_1^2\pi_1^*+\eta_2^2\pi_2^*)} > \frac{(\eta_1\pi_2^* + \eta_2\pi_1^*)^2}{\kappa(\eta_1^2+\eta_2^2)+(\eta_1^2\pi_2^*+\eta_2^2\pi_1^*)} \tag{40}$$

Expanding and using the fact that $\eta_1\pi_1^* + \eta_2\pi_2^* \geq \eta_1\pi_2^* + \eta_2\pi_1^*$, suffices to show:

$$(\eta_1^2\pi_2^* + \eta_2^2\pi_1^*)(\eta_1\pi_1^* + \eta_2\pi_2^*)^2 - (\eta_1^2\pi_1^* + \eta_2^2\pi_2^*)(\eta_1\pi_2^* + \eta_2\pi_1^*)^2 = \pi_1^*\pi_2^*(\pi_1^* - \pi_2^*)(\eta_1^2 - \eta_2^2)(\eta_1 - \eta_2)^2 > 0, \tag{41}$$

as desired. □

Proof of Proposition 4

Proof. By Lemma 3, suffices to assume without loss of generality that $(\gamma_1, \gamma_2) = (\bar{\gamma}, 0)$. Let (π_1, π_2) be the induced beliefs given that cue. One can easily see that the optimal portfolio given (π_1, π_2) is given by:

$$x^* = Var[R]^{-1}E[R] = \frac{1}{(\sigma^2 + \alpha^2\pi_1(1-\pi_1))(\sigma^2 + \alpha^2\pi_2(1-\pi_2)) - \alpha^4\pi_1^2\pi_2^2}\begin{bmatrix}\sigma^2\pi_1 + \alpha^2\pi_1\pi_2 \\ \sigma^2\pi_2 + \alpha^2\pi_1\pi_2\end{bmatrix}, \tag{42}$$

with the tilt of the portfolio given by

$$T(x^*) = \frac{\sigma^2\pi_1 + \alpha^2\pi_1\pi_2}{\sigma^2\pi_2 + \alpha^2\pi_1\pi_2}. \tag{43}$$

Denote for simplicity $\kappa \equiv \frac{\alpha^2}{\sigma^2}$. One can derive that $\frac{\partial T}{\partial \delta_1} < 0$ if and only if:

$$\pi_2(1 + \kappa\pi_2)\frac{\partial \pi_1}{\partial \delta_1} < \pi_1(1 + \kappa\pi_1)\frac{\partial \pi_2}{\partial \delta_1} \tag{44}$$

Recall that:

$$\pi_1 = \frac{\bar{\gamma} + \delta_0}{(1 + 2\delta_0) + \bar{\gamma}(1 + \delta_0 + \delta_1)},\ \pi_2 = \frac{\bar{\gamma}\delta_1 + \delta_0}{(1 + 2\delta_0) + \bar{\gamma}(1 + \delta_0 + \delta_1)}, \tag{45}$$

and in particular $\frac{\partial \pi_1}{\partial \delta_1} < 0$ and $\frac{\partial \pi_2}{\partial \delta_1} > 0$. This automatically implies the above inequality. □

Proposition 7. *The informative persuader locally moves beliefs equally for π_1 and π_2 and offers a perfectly diversified portfolio: $T(x^*_{inf}) = 1$.*

Proof. Recall that the cost of moving someone's beliefs (through information) from $(\pi_1^d, \pi_2^d, \pi_0^d)$ to (π_1, π_2, π_0) is given by $\xi \cdot D_{KL}(\pi \| \pi_d)$. Given our assumption that ξ is sufficiently high, suffices to consider marginal perturbations of beliefs near $(\pi_1^d, \pi_2^d, \pi_0^d)$.

We have from the above derivation that the optimal fees given beliefs π_1, π_2 is given by:

$$F(\pi_1, \pi_2) \equiv \frac{\alpha^2(\pi_1^2 + \pi_2^2)\sigma^2 + \alpha^2(\pi_1 + \pi_2)\pi_1\pi_2}{\sigma^4 + \alpha^2(\pi_1 + \pi_2)\sigma^2 + \pi_1\pi_2 \cdot \alpha^4 - \alpha^2\left[(\pi_1^2 + \pi_2^2)\sigma^2 + \alpha^2(\pi_1 + \pi_2)\pi_1\pi_2\right]}. \tag{46}$$

The informative persuader maximizes over local perturbations $\pi_1 = \pi_1^d + d\pi_1$ and $\pi_2 = \pi_2^d + d\pi_2$ the following expression:

$$\max F(\pi_1^d + d\pi_1, \pi_2^d + d\pi_2) - \xi\left[\log\left(\frac{\pi_1}{\pi_1^d}\right)\pi_1 + \log\left(\frac{\pi_2}{\pi_2^d}\right)\pi_2 + \log\left(\frac{\pi_0}{\pi_0^d}\right)\pi_0\right]. \tag{47}$$

The persuader objective admits the following Taylor approximation:

$$\nabla_1 F d\pi_1 + \nabla_2 F d\pi_2 - \frac{\xi}{2}\left(\frac{d\pi_1^2}{\pi_1^d} + \frac{d\pi_2^2}{\pi_2^d} + \frac{(d\pi_1 + d\pi_2)^2}{\pi_0^d}\right), \tag{48}$$

where $\nabla_1 F = \nabla_2 F \equiv \nabla > 0$. Putting this together, for sufficiently large ξ, the following FOC should hold:

$$0 = \nabla - \xi \cdot \frac{1}{\pi_1^d} d\pi_1 - \xi \frac{d\pi_1 + d\pi_2}{\pi_0^d} \implies d\pi_1^* = \frac{\nabla}{\xi} \cdot \left(\frac{1}{\pi_1^d} + \frac{1}{\pi_0^d}\right) = d\pi_2^*, \tag{49}$$

as desired. □

B.2 Consumer choice

Proof. We consider two extreme cases: $\gamma_2 = 0$ (complete separation) and $\gamma_2 = \bar{\gamma}_2$ (maximal association). In the first case, the induced beliefs (when a consumer considers good i) is given by:

$$(\pi_i, \pi_{-i}, \pi_0) = \frac{((\delta_0 + \gamma_1)p, (\delta_0 + \gamma_1\delta_1)p, (1 + \gamma_1\delta_0)p_0).}{(1 + \gamma_1\delta_0)p_0 + (2\delta_0 + \gamma_1(1 + \delta_1))p} \tag{50}$$

By symmetry, consumers that have greater weight for ω_i will select into the good that is specialized towards ω_i. Optimizing over the product design ψ net of quadratic costs, one

obtains the per-unit net profit of:

$$\Pi_{sep} = \frac{p^2\left(\eta^2(\delta_0+\gamma_1)^2+(\delta_0+\gamma_1\delta_1)^2\right)}{2C\cdot((1+\gamma\delta_0)p_0+(2\delta_0+\gamma(1+\delta_1))p)^2} \tag{51}$$

Analogously, the per-unit net profits for maximal association is given by:

$$\Pi_{pool} = \frac{\eta^2(\delta_0+\gamma_1+\bar{\gamma}_2\delta_1)^2p^2+(\delta_0+\gamma_1\delta_1+\bar{\gamma}_2)^2p^2}{2C\cdot((1+\gamma_1\delta_0)p_0+(2\delta_0+(\gamma_1+\bar{\gamma}_2)(1+\delta_1))p)^2} \tag{52}$$

Thus, $\Pi_{sep} > \Pi_{pool}$ if and only if:

$$\begin{aligned}
&\frac{\eta^2(\delta_0+\gamma_1)^2+(\delta_0+\gamma_1\delta_1)^2}{((1+\gamma_1\delta_0)p_0+(2\delta_0+\gamma_1(1+\delta_1))p)^2} > \frac{\eta^2(\delta_0+\gamma_1+\bar{\gamma}_2\delta_1)^2+(\delta_0+\gamma_1\delta_1+\bar{\gamma}_2)^2}{((1+\gamma_1\delta_0)p_0+(2\delta_0+(\gamma_1+\bar{\gamma}_2)(1+\delta_1))p)^2} \\
\iff &\frac{\eta^2(\delta_0+\gamma_1)^2+(\delta_0+\gamma_1\delta_1)^2}{\eta^2(\delta_0+\gamma_1+\bar{\gamma}_2\delta_1)^2+(\delta_0+\gamma_1\delta_1+\bar{\gamma}_2)^2} > \frac{((1+\gamma_1\delta_0)p_0+(2\delta_0+\gamma_1(1+\delta_1))p)^2}{((1+\gamma_1\delta_0)p_0+(2\delta_0+(\gamma_1+\bar{\gamma}_2)(1+\delta_1))p)^2}
\end{aligned} \tag{53}$$

As stated in the Proposition, we assume for simplicity that $\bar{\gamma}_2$ is small relative to 1 and γ_1 (the effect of cross-product associations is second order relative to the default cue and the direct advertising campaign). In that case, the LHS and RHS simplifies to:

$$1-2\cdot\frac{(1+\delta_1\eta^2)\delta_0+(1+\eta^2)\delta_1\gamma_1}{\eta^2(\delta_0+\gamma_1)^2+(\delta_0+\gamma_1\delta_1)^2}\cdot\gamma_2+O(\gamma^2) > 1-2\cdot\frac{(1+\delta_1)p}{(1+\gamma_1\delta_0)p_0+(2\delta_0+\gamma_1(1+\delta_1))p}\cdot\gamma_2+O(\gamma^2) \tag{54}$$

Setting $\tilde{p} \equiv p_0/p$, $\Pi_{sep} > \Pi_{pool}$ if and only if:

$$\frac{(1+\delta_1)}{(1+\gamma_1\delta_0)\tilde{p}+(2\delta_0+\gamma_1(1+\delta_1))} > \frac{(1+\delta_1\eta^2)\delta_0+(1+\eta^2)\delta_1\gamma_1}{\eta^2(\delta_0+\gamma_1)^2+(\delta_0+\gamma_1\delta_1)^2} \tag{55}$$

Finally, as assumed, we exploit $\delta_0 \ll \delta_1 < 1$, which simplifies the above into showing:

$$\begin{aligned}
\frac{1+\delta_1}{\tilde{p}+\gamma_1(1+\delta_1)} > \frac{(1+\eta^2)\delta_1\gamma_1}{(\eta^2+\delta_1^2)\gamma_1^2} &\iff \gamma_1^2(\delta_1^2+\eta^2)(1+\delta_1)-(1+\eta^2)\delta_1\gamma_1^2(1+\delta_1) > (1+\eta^2)\gamma_1\tilde{p}\delta_1 \\
&\iff \gamma_1^2(1+\delta_1)(\eta^2-\delta_1)(1-\delta_1) > (1+\eta^2)\gamma_1\tilde{p}\delta_1
\end{aligned} \tag{56}$$

Note that the LHS is a cubic with a positive leading coefficient in δ_1 with intercepts in -1, 1, and $\eta^2 > 1$, and the RHS is positive and linear in δ_1. This implies that for $\delta \in (0,1)$, there is a single crossing between the LHS and the RHS at $\delta = \delta^* \in (0,1)$, with the LHS

being greater than the RHS for $\delta \in (0, \delta^*)$ and less than the RHS for $\delta \in (\delta^*, 1)$, as desired.

The result regarding η also holds immediately from the fact that the RHS

$$\frac{(1+\eta^2)\delta_1\gamma_1}{(\eta^2+\delta_1^2)\gamma_1^2} \tag{57}$$

is an increasing function in η from the fact that $\delta_1 < 1$.

□

C Application to political persuasion

We apply our framework to the setting of political persuasion. Any public policy (i.e. welfare, education bill, etc) entails a complex transfer from one group of agents to another. In that case, by highlighting how a policy effects a particular group of agents, a persuader is able to influence the public support for a given policy.

Setup We define a policy by how it affects each agent ω in the economy, where the net benefit of the policy for ω is given by $NB(\omega)$. An agent ω's utility given the policy Ψ is given by:

$$U_{rat}(\omega|\Psi) = NB(\omega|\Psi) + \eta \cdot \sum_{\omega' \in \Omega} NB(\omega'|\Psi) \cdot \lambda(\omega'|\omega) - C(\Psi), \tag{58}$$

where $\lambda(\omega'|\omega)$ is the subjective Pareto weight that agent ω places on ω', which reflects the social preferences of ω and $C(\Psi)$ are residual costs or benefits associated with policy Ψ. As before, the persuader highlights a subset of agents $Q \subset \Omega$, which influences the strength of simulation of ω', $\pi(\omega'|\omega) \propto S(\omega', d) + \gamma \cdot S(\omega', Q)$. One can derive the following formula for ω's support for a policy Ψ given $C = (d, Q)$:

$$U(\omega|Q) \equiv \sum_{\omega' \in \Omega} NB(\omega'|\Psi) \cdot \lambda(\omega'|\omega) \cdot \pi(\omega'|\omega) - C(\Psi) = Sup_{rat}(\omega) + \frac{Cov\left[\lambda(\omega) \cdot NB(\omega), S_C(\omega)\right]}{A(d) + \gamma A(Q)}, \tag{59}$$

where $S_C(\omega) = S(\omega, d) + \gamma S(\omega, Q)$. A persuader can bolster can support for a policy by disproportionately cuing those that benefit from the policy that agents want to help.

Application: redistribution preferences and out-groups Consider a politician who wishes to persuade the population to adopt a less redistributive policy. As equation 59 indicates, the politician will do so by highlighting a beneficiary whom the majority thinks is undeserving (a low λ). A particularly salient case is where the politician can exploit ma-

jority distaste of an "out-group" (a minority for which the majority bears a lower Pareto weight) – e.g. immigrants or ethnic minorities – to undermine support for redistributive policies. Famous cases include the welfare queen, the racialization of redistributive politics and the issue of immigration (Alesina et al., 2023). We model this phenomenon directly. Not only do we find that such a strategy can undermine the support for redistribution, but also has broader impact on how the out-group is perceived, with potentially pernicious welfare impact, especially for the majority poor.

Let each agent ω have income x_ω, with the redistributive policy restricted to a proportional tax with a flat rebate (Alesina & Giuliano, 2011; Hvidberg et al., 2023):

$$NB(\omega) \equiv c_\omega = (1-\tau)x_\omega + \frac{1}{|\Omega|}\tau \sum_\omega x_\omega. \tag{60}$$

Agent ω's preferences over τ is shaped by how much she simulates each agent ω', $\pi(\omega'|\omega)$. To ensure an interior τ^*, we further assume relative to equation 58 that the agent internalizes the distortive cost of the tax ($\frac{1}{2}\frac{1}{\phi}\tau^2 x_{\omega'}$ for each $\omega' \in \Omega$):

$$U(\tau|\omega, d, Q) \equiv (1-\tau)x_\omega + \tau\bar{x}_\pi + \eta \sum_{\omega'}((1-\tau)x_{\omega'} + \tau\bar{x}_\pi)\lambda(\omega'|\omega)\pi(\omega'|\omega) - \frac{1}{2}\frac{1}{\phi}\sum_{\omega'}\tau^2\pi(\omega'|\omega)x_{\omega'}, \tag{61}$$

where $\bar{x}_\pi \equiv \sum_{\omega'}\pi(\omega'|\omega)x_{\omega'}$.

To explicitly solve for the model, we specify the associative structure $S(\omega, \omega')$ as well as the social preferences $\lambda(\omega'|\omega)$ of all agents in the economy. There are two groups of agents: the majority T_M, who forms a mass $1-q$ of the total population, and the minority T_m, who forms the remaining mass q. The similarity between ω and ω' is determined by their relative income difference as well as whether they belong to the same group:

$$S(\omega, \omega') = \delta^{I(T_\omega = T_{\omega'})} \cdot \exp\left(-\frac{\kappa}{2}(x_\omega - x_{\omega'})^2\right). \tag{62}$$

For social preferences, we assume across-group enmity: $\lambda(\omega'|\omega) = 1$ if $T_\omega = T_{\omega'}$ and $-\lambda_0 < 0$ otherwise.[12] Finally, assume that the default context $d(\omega)$ for agent ω is given by an agent of modal income μ of her group: $d = (T_\omega, \mu)$. Let us focus on the simple case where the objective distribution of income across these groups is identical: $x_\omega \sim N(\mu, \sigma^2)$ for both groups.

[12] In particular, for the sake of simplicity we are shutting down for explicit fairness motives (with lower λ for richer agents) that traditionally motivate preferences for redistribution. This is to clearly illustrate the role of highlighting outgroups in distorting redistributive preferences. Similar conclusions hold when one generalizes λ to account for fairness motives.

Proposition 8 (Political persuasion and redistributive preferences)**.** *Suppose that a politician highlights a minority group with lower income,* (T_m, x^*)*, where* $x^* < \mu$*. Every agent in the economy perceives a majority-minority income gap: the majority group is perceived as having higher income than the minority group. Assume for simplicity that the equilibrium redistributive policy* τ^* *settles to the median majority preference.*[13] *Equilibrium policy is less redistributive than the persuader-less benchmark* $(\tau^* < \tau_d^*)$*, with the policy change disproportionately hurting the poor of both groups.*

The key intuition behind Proposition 8 is that a persuader can shape redistributive policies by influencing how an agent represents and simulates the poor. Under $Q = (T_m, x^*)$, voters increasingly imagine a poor minority as the primary beneficiary of welfare policies. Such a strategy undermines the support for redistributive policies in two ways. First, by highlighting the poor, the persuader lowers the perceived distribution of income in the economy: the median voter feels "relatively richer" as a result of the cue and thinks a given redistributive policy would lower their income. Second, by highlighting a minority poor, the persuader creates the illusion of across-group income disparity (or in the case where there is actual across-group disparity, worsens the perception of the income gap). This implies that a redistributive policy not only helps the poor, but disproportionately helps the minority, which is politically less popular.

This type of persuasion has two potentially pernicious effects, for both majority and the minority. First, the biggest losers of the policy shift are ironically the majority poor. While they believe they are trading off lower post-tax income with "punishing" the minority group, the motivating across-group differences are exaggerated (and purely illusory in this simple case).[14] Second, the strategy results in exacerbating the negative stereotypes regarding minority members. In other words, there can be a scope for possibly pernicious long-term effects of political rhetoric and cuing that is effective for short-term goals: while the rhetoric of the welfare queen was effective in eroding support for redistributive politics, the racial animus and stereotypes it triggered has much long-run impact on social cohesion and democratic politics.

[13] The assumption is made for simplicity and not strictly necessary. A similar conclusion holds when the policy settles to the total median preferred τ^*, provided that the relative fraction of minorities is sufficiently small.

[14] Some political references helpful here: a characterization of anti-welfare front being united between the rich and the majority poor?

C.1 Proofs

Suppose that the politician highlights a hypothetical minority agent with income x^*. Then, for the majority group, the context C is given by the default (T_M, μ) and the persuader provided (T_m, x^*), which has weight γ. These cues in turn distort the implicit distribution of agents in the economy. Here is a following useful lemma:

Lemma 4. *Given the similarity function* $S(\omega, \omega') = \exp\left(-\frac{\kappa(\omega-\omega')^2}{2}\right)$ *and a normal distribution* $\phi_{\mathcal{N}}(\mu, \sigma^2)$*, the similarity-distorted distribution* $S(\omega, \omega') \cdot \phi_{\mathcal{N}}(\mu, \sigma^2)$ *is exactly given by* $\xi((\mu - x^*)^2, \kappa, \sigma^2) \cdot \phi_{\mathcal{N}}\left(\frac{\mu+\kappa\sigma^2 x^*}{1+\kappa\sigma^2}, \frac{\sigma^2}{1+\kappa\sigma^2}\right)$*, where* $\xi((\mu - x^*)^2, \kappa, \sigma^2) \equiv \frac{1}{\sqrt{1+\kappa\sigma^2}} \cdot \exp\left(-\frac{\kappa(\mu-x^*)^2}{2(1+\kappa\sigma^2)}\right)$.

Proof. The expression follows from the following standard algebraic manipulation:

$$\begin{aligned}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2} - \frac{\kappa(x-x^*)^2}{2}\right) &= \exp\left(-\frac{\kappa(\mu-x^*)^2}{2(1+\kappa\sigma^2)}\right)\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(1+\kappa\sigma^2)}{2\sigma^2}\left(x - \frac{\mu+\kappa\sigma^2 x^*}{1+\kappa\sigma^2}\right)^2\right) \\ &= \xi((\mu-x^*)^2, \kappa, \sigma^2) \cdot \phi_{\mathcal{N}}\left(\frac{\mu+\kappa\sigma^2 x^*}{1+\kappa\sigma^2}, \frac{\sigma^2}{1+\kappa\sigma^2}\right).\end{aligned} \tag{63}$$

□

Applying the lemma, note that under the cue, the income distribution of non-minorities get distorted into (up to a proportionality constant):

$$(1-q) \cdot \left[\xi(0, \kappa, \sigma^2) \cdot \phi_{\mathcal{N}}\left(\mu, \frac{\sigma^2}{1+\kappa\sigma^2}\right) + \delta\gamma \cdot \xi((\mu-x^*)^2, \kappa, \sigma^2) \cdot \phi_{\mathcal{N}}\left(\frac{\mu+\kappa\sigma^2 x^*}{1+\kappa\sigma^2}, \frac{\sigma^2}{1+\kappa\sigma^2}\right)\right], \tag{64}$$

while the income distribution of minorities get distorted to:

$$q \cdot \left[\delta \cdot \xi(0, \kappa, \sigma^2) \cdot \phi_{\mathcal{N}}\left(\mu, \frac{\sigma^2}{1+\kappa\sigma^2}\right) + \gamma \cdot \xi((\mu-x^*)^2, \kappa, \sigma^2) \cdot \phi_{\mathcal{N}}\left(\frac{\mu+\kappa\sigma^2 x^*}{1+\kappa\sigma^2}, \frac{\sigma^2}{1+\kappa\sigma^2}\right)\right]. \tag{65}$$

Thus, contextual distortions imply that the majority perceives the economy as consisting of:

$$w_M^* \equiv \frac{(1-q) \cdot (\xi(0, \kappa, \sigma^2) + \delta\gamma \cdot \xi((\mu-x^*)^2, \kappa, \sigma^2))}{q \cdot (\delta \cdot \xi(0, \kappa, \sigma^2) + \gamma \cdot \xi((\mu-x^*)^2, \kappa, \sigma^2)) + (1-q) \cdot (\xi(0, \kappa, \sigma^2) + \delta \cdot \gamma \cdot \xi((\mu-x^*)^2, \kappa, \sigma^2))} \tag{66}$$

majority and $w_m^* = 1 - w_M^*$ minority agents. The conditional mean income of each group

is given by:

$$\begin{aligned}\mu_M &\equiv \mu - \frac{\delta\gamma\cdot\xi((\mu-x^*)^2,\kappa,\sigma^2)}{\xi(0,\kappa,\sigma^2)+\delta\gamma\cdot\xi((\mu-x^*)^2,\kappa,\sigma^2)}\cdot\frac{\kappa\sigma^2}{1+\kappa\sigma^2}(\mu-x^*)\\ \mu_m &\equiv \mu - \frac{\gamma\cdot\xi((\mu-x^*)^2,\kappa,\sigma^2)}{\delta\cdot\xi(0,\kappa,\sigma^2)+\gamma\cdot\xi((\mu-x^*)^2,\kappa,\sigma^2)}\cdot\frac{\kappa\sigma^2}{1+\kappa\sigma^2}(\mu-x^*)\end{aligned} \tag{67}$$

For $x^* < \mu$, both μ_M and μ_m are negative – highlighting any poor agent tends to raise the availability of all poor agents. However, note that due to $\delta < 1$, one immediately obtains $\frac{\gamma\cdot\xi((\mu-x^*)^2,\kappa,\sigma^2)}{\delta\cdot\xi(0,\kappa,\sigma^2)+\gamma\cdot\xi((\mu-x^*)^2,\kappa,\sigma^2)} > \frac{\delta\gamma\cdot\xi((\mu-x^*)^2,\kappa,\sigma^2)}{\xi(0,\kappa,\sigma^2)+\delta\gamma\cdot\xi((\mu-x^*)^2,\kappa,\sigma^2)} \implies \mu_M \geq \mu_m$: there is an induced perceived income distortion between majority and minority agents. Intuitively, highlighting a poor minority disproportionately lowers the perceived income of the minority group.

How does this distortion in turn impact the preference for redistribution? The first-order condition implies:

$$\begin{aligned}\tau^*(\omega) &= \operatorname*{argmax}_{\tau}(1-\tau)x_\omega + \tau\bar{x}_\pi + \eta\sum_{\omega'}((1-\tau)x_{\omega'}+\tau\bar{x}_\pi)\lambda(\omega'|\omega)\pi(\omega'|\omega) - \frac{1}{2}\frac{1}{\phi}\sum_{\omega'}\tau^2\pi(\omega'|\omega)x_{\omega'}\\ &= -\phi\left(\tilde{x}_\omega + \eta\sum_{\omega'}\lambda(\omega'|\omega)\pi(\omega'|\omega)\cdot\tilde{x}_{\omega'}\right),\end{aligned} \tag{68}$$

where $\tilde{x}_\omega = \frac{x_\omega-\bar{x}_\pi}{\bar{x}_\pi}$. Note that $\bar{x}_\pi = w_M^*\cdot\mu_M + w_m^*\cdot\mu_m$, and the social preferences then are given by:

$$\begin{aligned}\tau^*(\omega) &= -\frac{\phi}{\bar{x}_\pi}\cdot\left[x_\omega - \bar{x}_\pi + \eta\left(w_M^*(\mu_M-\bar{x}_\pi) - \lambda_0 w_m^*(\mu_m-\bar{x}_\pi)\right)\right]\\ &= -\frac{\phi}{\bar{x}_\pi}\cdot\left[x_\omega - \bar{x}_\pi + \eta(1+\lambda_0)w_M^*w_m^*(\mu_M-\mu_m)\right]\end{aligned} \tag{69}$$

Assuming that the policy settles to the median preferences. In that case, we have:

$$\tau^* = -\frac{\phi}{\bar{x}_\pi}\cdot\left[\mu - \bar{x}_\pi + \eta(1+\lambda_0)w_M^*w_m^*(\mu_M-\mu_m)\right]. \tag{70}$$

In contrast, consider the case where the context is given by the default context (in the absence of the persuader). In that case, there is no distortion in the relative income of both populations, with the redistributive preferences given by:

$$\tau^*(\omega)_d = -\frac{\phi}{\mu}\cdot[x_\omega - \mu], \tag{71}$$

with the median preferences given by:

$$\tau_d^* = 0. \tag{72}$$

Denote $\mu_M = \mu - \alpha_M \Delta$, $\mu_m = \mu - \alpha_m \Delta$, where $\alpha_m > \alpha_M$. We have:

$$\bar{x}_\pi = \mu - (w_M^* \alpha_M + w_m^* \alpha_m)\Delta, \tag{73}$$

so

$$\tau^* = -\frac{\phi}{\bar{x}_\pi}\left[(w_M^* \alpha_M + w_m^* \alpha_m) + \eta(1+\lambda_0) w_M^* w_m^* (\alpha_m - \alpha_M)\right]\Delta < \tau_d^* \tag{74}$$

Finally, welfare consequences: let's compare the objective welfare difference from two redistributive policies τ_d^* and τ^*:

$$U(\tau^*) - U(\tau_d^*) = (\tau_d^* - \tau)^*(x_\omega - \mu) - \frac{1}{2}\frac{1}{\phi}((\tau^*)^2 - (\tau_d^*)^2)\mu. \tag{75}$$

As the expression makes clear, a less redistributive policy ($\tau^* < \tau^* d$) directly hurts agents with lower income in both in-group and the out-group. Critically, the majority poor think that they are accepting a lower income to avoid an unwanted net transfer to the minority. In the simple example where *there is no across-group inequality*, in reality there is no such across-group transfers. Hence, only the direct impact of the redistributive policy plays a role in the objective welfare.

In a more general case where there is actual group income inequality – the minority members are poorer than the majority – the majority members objectively benefit from having lower redistribution (because it results in less subsidies to the outgroup). However, the political persuasion over-accentuates this benefit, resulting in a similar conclusion: the majority poor are overpaying to punish the minority.