Choice with Competing Models: An Experimental Study

Sandro Ambuehl ® Heidi C. Thysen *

November 13, 2025

Abstract

People often make choices with imperfect knowledge of how the variables in their decision problem are related. We study such choices when individuals face menus of conflicting and possibly misspecified models that link these variables. Do they discard inaccurate models, what types of inaccuracies do they detect, and how? Or do they instead follow models that sound appealing at face value, and what determines that appeal? Our experiment yields two main findings. First, many individuals readily intuit the models' predicted correlations and reject models that contradict the data. Performance is high because the required inference is qualitative rather than quantitative. Second, when unable to identify the correct model, most choose cautiously by focusing on worst-case outcomes. This behavior contradicts the Narrative Competition literature's assumption of best-case maximization, but a failure of contingent reasoning when interpreting models' payoff implications can mimic that assumption. Our results are robust to tripled stakes.

JEL codes: C91, D01, D83

*Ambuehl: Department of Economics and UBS Center for Economics in Society, University of Zurich, Blüemlisalpstrasse 10, 8006 Zürich, Switzerland, sandro.ambuehl@econ.uzh.ch. Thysen: Department of Economics, Norwegian School of Economics, Helleveien 30, 5045 Bergen, Norway, heidi.thysen@gmail.com. A previous version of this draft based on a different design and different experimental data circulated under the title 'Competing Causal Interpretations: An Experimental Study.' We are grateful to Chiara Aina, Gonzalo Arrieta, Björn Bartling, Ernst Fehr, Gilat Levy, Chad Kendall, Nick Netzer, Rani Spiegler, Jakub Steiner, Severine Toussaert, Bertil Tungodden, Roberto Weber, and Georg Weizsäcker for helpful comments and suggestions, to Timo Huber for his excellent programming of the experiment, and to Eva Küpper and Nandita Gupta for excellent research assistance. This project was funded by a generous grant from FAIR – Centre for Experimental Research on Fairness, Inequality and Rationality at the Norwegian School of Economics, the Department of Economics at the University of Zurich, and the SNSF Starting Grant #211222. All experiments were approved by the Ethics Review Board of the Department of Economics at the University of Zurich and the Norwegian School of Economics.

1 Introduction

A vast literature on belief updating in economics studies how individuals use information for choice when the structure of the data-generating process (DGP) is known (see, e.g. Benjamin, 2019). Yet, many decision problems lack structural information, such as which variables in the decision problem are related and in what way. How do individuals learn and choose when they face uncertainty about the structure of the data-generating process?

We consider a setting in which individuals face a menu of conflicting and possibly misspecified models fitted to the DGP. These models make claims about the relationships between variables and differ in their implications for choice and utility. Individuals decide which model to follow. This framework captures, for instance, political referenda in which voters choose among a small number of proposals, each based on a different structural understanding of the world. It likewise applies to managers choosing among a few action plans prepared by teams with differing structural views of the competitive environment.

We study how individuals choose between such models. On the one hand, individuals may recognize and discard models that are inconsistent with the data (*Fact-Based* decision making). If so, how do they make these decisions? Do they look for qualitative inconsistencies or do they focus on broad quantitative fit? What model implications can they intuit and connect to the data, and what types of inconsistencies do they fail to recognize? On the other hand, individuals may follow models whose implications for utility and choice seem appealing at face value (*Utility-Based* decision making). If so, what determines that appeal? Is it the utility a model promises under the assumption that it is correct—which is the assumption that drives the literature on Narrative Competition (Eliaz and Spiegler, 2020; Eliaz et al., 2022; Levy et al., 2022)? Or do they instead favor caution, as suggested by the literature on choice under uncertainty (see, e.g., Trautmann and Van De Kuilen, 2015; Gilboa, 2025)?

We study these questions in a laboratory experiment. This method provides full control over the data-generating process and it allows us to abstract from confounding factors such as prior beliefs or attachment to political groups.

Our setting involves four variables: the investment amount (which the subject chooses), the payoff amount, and two additional variables we call *covariates*. These variables are causally linked through a data-generating process (DGP) that the subject does not know. Instead, the subject observes two competing models that we convey in natural language, using terms such as 'X directly affects Y,' 'X influences Y indirectly through Z,' or 'X is a symptom of Z,' along with graphical illustrations. Models differ in the causal roles they assign to the four variables, possibly erroneously. While each model treats the investment as exogenous, the payoff and the covariates might be endogenous or exogenous. They might be causes or symptoms of other variables, or mediators between them. We fit each model to data produced by the DGP. Subjects observe the fitted model's implications for choice and utility. They have access to (simulated) empirical data from the DGP which they may use to rule out misspecified models. After studying the models, their implications, and the empirical data, subjects select between two investment levels, each implied as optimal by one of the models (but possibly suboptimal according to the DGP), and have their payoff determined according to their choice and the DGP.

Subjects make a choice in each of several rounds. Each round features a different DGP and a different pair of models fit to it, along with different implications for the optimal investment levels, as well as different claims about best- and worst-case outcomes. We formulate our hypotheses as a list of decision criteria subjects might use. We systematically vary the available models and their implications across the rounds to ensure that each combination of decision criteria corresponds to a unique pattern of choices across all rounds. This revealed preference information lets us identify subjects' decision making. Formally, we estimate a finite mixture model to identify the frequency of each decision criterion.

We document two main results. First, subjects display a remarkable ability to discard misspecified models by connecting models and data. Three fifths make purely Fact-Based choices. They not only intuit the relevant correlational implications but also find which of 18 possible charts displaying empirical data will help them check these implications. These subjects show no tendency to detect some types of misspecification more often than others. An additional fifth detect only misspecifications revealed by unconditional, but not by conditional, correlational

data. Half of the latter succeed only when the model-inconsistent correlation is implied by a proposed direct causal effect, but not when it results from an indirect effect or a common cause affecting two variables. The final fifth of subjects do not make any consistent use of empirical information. A treatment that raises the stakes by a factor of three (up to \$90 in some cases) does not affect these results. Hence, failures to detect certain inconsistencies reflect limited motivation rather than limited ability.

Our second main result is that subjects unable or unwilling to rule out misspecified models choose cautiously. In our main treatment, roughly one fifth of subjects ever deploy Utility-Based criteria. Caution describes the vast majority of these choices, whereas best-case maximizing choices—the assumption of the Narrative Competition literature—are extremely rare. An additional treatment restricts access to model structures and empirical data to reveal the Utility-Based criteria of the remaining four fifths of subjects. In this treatment, two thirds choose cautiously, while the remaining third maximizes best-case outcomes. This result arises when we communicate a model's implications as its prediction about the maximally achievable utility, along with its prediction about the utility expected if the competitor's recommendation is followed. While this presentation is natural for models fit to data, it differs from the presentation of payoff information in typical risky-choice experiments (which is unnatural in our setting). In treatments that present payoff implications in the latter way, the fraction of best-case maximizing subjects drops to a tenth. A treatment involving dominated choices shows that the natural way of presenting implications of models fit to data induces errors, likely due to a failure of contingent thinking (Niederle and Vespa, 2023), and hence does not reflect actual preferences.

Our subjects' ability to rule out incorrect models may seem surprising given the literature on belief updating (Benjamin, 2019), especially since we provide no assistance. Our data suggests that their performance is due to the possibility to rely on qualitative rather than quantitative inference. In principle, subjects could assess a model's correctness through quantitative reasoning, as each round provides data on the relation between investment and outcome according to the DGP, from which they could infer the optimal investment. Yet few subjects view that data, and virtually none view it exclusively. Instead, they intuit the mod-

els' correlational implications and check the empirical correlations on which the implications differ. They rarely view empirical correlations on which the models' qualitative implications coincide. Rounds in which both models are misspecified provide further evidence for our hypothesis of qualitative rather than quantitative inference: subjects tend to pick the model with fewer inconsistencies between model-implied and empirical correlations rather than the one whose implied investment is closer to the empirically optimal level (which they could eyeball from the empirical data).

Out-of-sample predictions demonstrate that our estimates capture stable behavioral tendencies. They also reveal that a three-type model largely suffices to explain the bulk of predictable variation in the data. That model includes (i) a type that rules out all misspecified models, (ii) a type that excludes misspecified models only if a proposed direct causal effect appears absent in the empirical data (and otherwise randomizes), and (iii) a type that does not seek to or fails to discard misspecified models but makes choices to maximize worst-case payoffs.

Concerns that we may have enlisted unrepresentatively sophisticated subjects are unwarranted. While we observe the expected relationships between subjects' choices, their educational background, and their Cognitive Reflection Test scores (Frederick, 2005; Thomson and Oppenheimer, 2016), our subjects' statistical background is limited. Moreover, their CRT performance mirrors that of other university subject pools and of financial professionals. Other subject characteristics correlate only weakly or not at all with choices in our setting. Belief in pseudoscience (Torres et al., 2020) is directionally but statistically insigificantly associated with worse decision making, and neither political position nor political centrism have significant predictive power. This result contrasts with the common view that disagreements with one's own political views stem from others' objective inferential errors, but that view itself may be mistaken (*naïve realism*, Griffin and Ross, 1991)

Broadly, our results provide empirical foundations for the emergent literature on mental models whose applications span fields as diverse as behavioral economics (Spiegler, 2016), macroeconomics (Molavi, 2019), finance (Molavi et al., 2021; Shiller, 2017), strategic management (Felin and Zenger, 2017; Camuffo et al.,

2023), institutional economics (Denzau and North, 1994), and contract theory (Schumacher and Thysen, 2022). It makes three specific main contributions.[1]

First, it advances the literature on how individuals draw inferences from data (reviewed in Benjamin, 2019) by studying the case in which individuals lack a-priori information about the structure of the data-generating process. While a large literature in cognitive science (see, e.g., Waldmann, 2017; Sloman, 2005; Griffiths et al., 2024, for reviews) also addresses that question, that literature focuses on learning from scratch.[2] It finds that humans are generally much more adept in learning about structure than in updating proabilistic beliefs in the types of experiments reviewed in Benjamin (2019). Our paper's focus—choice between competing candidate models—permits decision strategies that differ fundamentally from those suited to learning from scratch. In particular, it allows for *constraint-based learning* (Spirtes et al., 2000), which tests whether the data satisfy the correlational constraints implied by a given model.

Second, our results inform the literature that interprets misspecified models as 'lens through which people view the world' (e.g., Schwartzstein and Sunderam, 2021; Kendall and Charles, 2022; Andre et al., 2023). We show that when models are provided externally, individuals are unlikely to blindly accept any such lens. Instead, they are adept at dismissing misspecified models when the requisite empirical data is available and the inconsistencies between models and data are qualitative rather than quantitative.

The third main contribution of our work is to test the main behavioral assumption of the Narrative Competition literature: decision makers adopt whichever

[1] A previous version of this paper reports the results from an earlier experiment. (Subjects who had participated in that experiment could not participate in the present experiment.) That experiment also found a remarkable ability of subjects to discard misspecified models, and a preference for cautious choice when subjects could not exclude misspecified models. It also found around 15% of subjects making choices consistent with the Best-Case Promise criterion. Its design has several shortcomings relative to the current experiment. First, it did not present information about the payoffs from following a competing model, which constitutes a possible reason for Best-Case Promise choice. Second, it did not include anything like our Utility-Focused part. Third, it artificially simplified the problem for subjects by highlighting the data charts for which the two models made differing predictions and by providing some explanation about the correlational implications of causal structures.

[2] A narrow exception, Steyvers et al. (2003) test whether subjects can distinguish the collider $X \rightarrow Y \leftarrow Z$ from the fork $X \leftarrow Y \rightarrow Z$ using a sequence of individual observations. In contrast to our study, Steyvers et al. (2003) show subjects trial-by-trial data. Hence, in their work, subpar performance may reflect an inability to extract correlational information from such data rather than an inability to recognize inconsistencies between predicted and empirical correlations.

model promises the highest utility *if correct*. That literature derives fascinating implications of such behavior—mutually inconsistent narratives will necessarily coexist; it is possible to predict which narratives will survive competition and which ones will not (Eliaz and Spiegler, 2020);[3] and cycles of populism will emerge (Levy et al., 2022).[4] Our data show that a substantial minority of individuals will engage in such Best-Case maximization due an interaction of the natural mode of communicating model implications with a failure of contingent reasoning (Niederle and Vespa, 2023), but only as long as they cannot access data to discard misspecified models.

Several recent studies in economics relate to our work. First, about three fifths of the subjects in Frechette et al. (2023) can learn and reproduce the structure of three-variable Bayesian networks from a list of observations. While their subjects' remarkable performance parallels ours', that work does not involve externally proposed, potentially misspecified models, and hence fundamentally differs from our tests of whether people can recognize inconsistencies between models and data. Second, Kendall and Charles (2022) study the effect of providing chain or collider narratives to subjects required to interpret a line-by-line dataset of three binary variables. In a setting with stakes of a few cents,[5] they find that providing subjects with a single narrative affects their choices in the direction consistent with that narrative, as do Barron and Fries (2023) in a different setting. They also find that subjects presented with two conflicting narratives choose intermediate actions. Despite the ostensible similarity to our work, their research questions

[3]For example, consider the question of whether mask-wearing causally reduces COVID-19 transmission. Focus on two causal models. Model 1 accurately describes reality and states that increased masking reduces COVID-19 transmission. Model 2 surmises that masking has no effect on disease transmission. If most individuals adopt the first model, they will wear masks, and case counts will be low. According to Model 1, this situation can be maintained only by continued masking, which has small hassle costs. Model 2 is more attractive because it predicts that ending mask-wearing will eliminate hassle costs without affecting case counts. The literature assumes that due to its greater attractiveness, individuals will flock to Model 2. Contrarily, if Model 2 is more popular, individuals will not wear masks, and case counts will be high. According to Model 2, masking cannot change this situation. Model 1 makes the more attractive prediction that case counts can be lowered at a small hassle cost. Individuals will thus flock to Model 1. Overall, the more popular one model, the more attractive the other. Accordingly, no model, including the correct one, can survive alone; multiplicity is an equilibrium.

[4]The formal assumption in Levy et al. (2022) is that the difference in the maximal utilities promised by two parties motivates turnout.

[5]A subject in their main sample who makes the worst choices loses 33 cents relative to optimal choice.

fundamentally differ from ours. We are interested in settings such as voting on political referenda or selecting between leaders with differing views of the interactions between economic variables. In these settings, agents must choose between models, and intermediate choices are unavailable. More importantly, our work generally characterizes the limits of subjects' inferential abilities in a way that is not constrained to chain or collider narratives, but rather concerns a large set of models and DGPs. Additionally, we study the prevalence and nature of Utility-Based criteria. Third, Angrisani et al. (2023) find that the evolution of beliefs during the Covid19 pandemic is consistent with the model of Eliaz and Spiegler (2020). While their structural model fits the data well, their field data naturally limits the extent to which they can rule out alternative explanations. More distantly related, Aina and Schneider (2025) ask how subjects quantitatively update beliefs in a balls-and-urns setting when each of two competing information structures could have generated a given signal. Their setting does not allow discarding information structures based on inconsistencies with the data.

The remainder of this paper proceeds as follows. Section 2 outlines the choice setting and defines the choice criteria we study. Section 3 explains our identification strategies along with details concerning the experimental design. Section 4 showcases our main empirical results. Finally, Section 5 concludes.

2 Setting and choice criteria

2.1 Choice problem

In each round of our experiment, subjects choose between two investment levels I. They know that the investment maps into an outcome Y, and that their payoff will be $\pi(I) = Y - I$, but they do not know the data-generating process (DGP) that determines whether and how the investment affects the outcome. DGPs involve four variables,[6] the investment I, the outcome Y, and two covariates C_1 and C_2, represented to subjects, respectively, as  ,  , and two counters of different colors (). These variables are related through a recursive system of Gaussian ('regression') equations that are linear in C_1, C_2, Y, and in the square

[6]This is the minimum number of variables that allows us to answer our research questions.

root of I. Investment I is always exogenous, all other variables may be endogenous or exogenous. Subjects simply learn that the variables interact through some 'mechanism.'

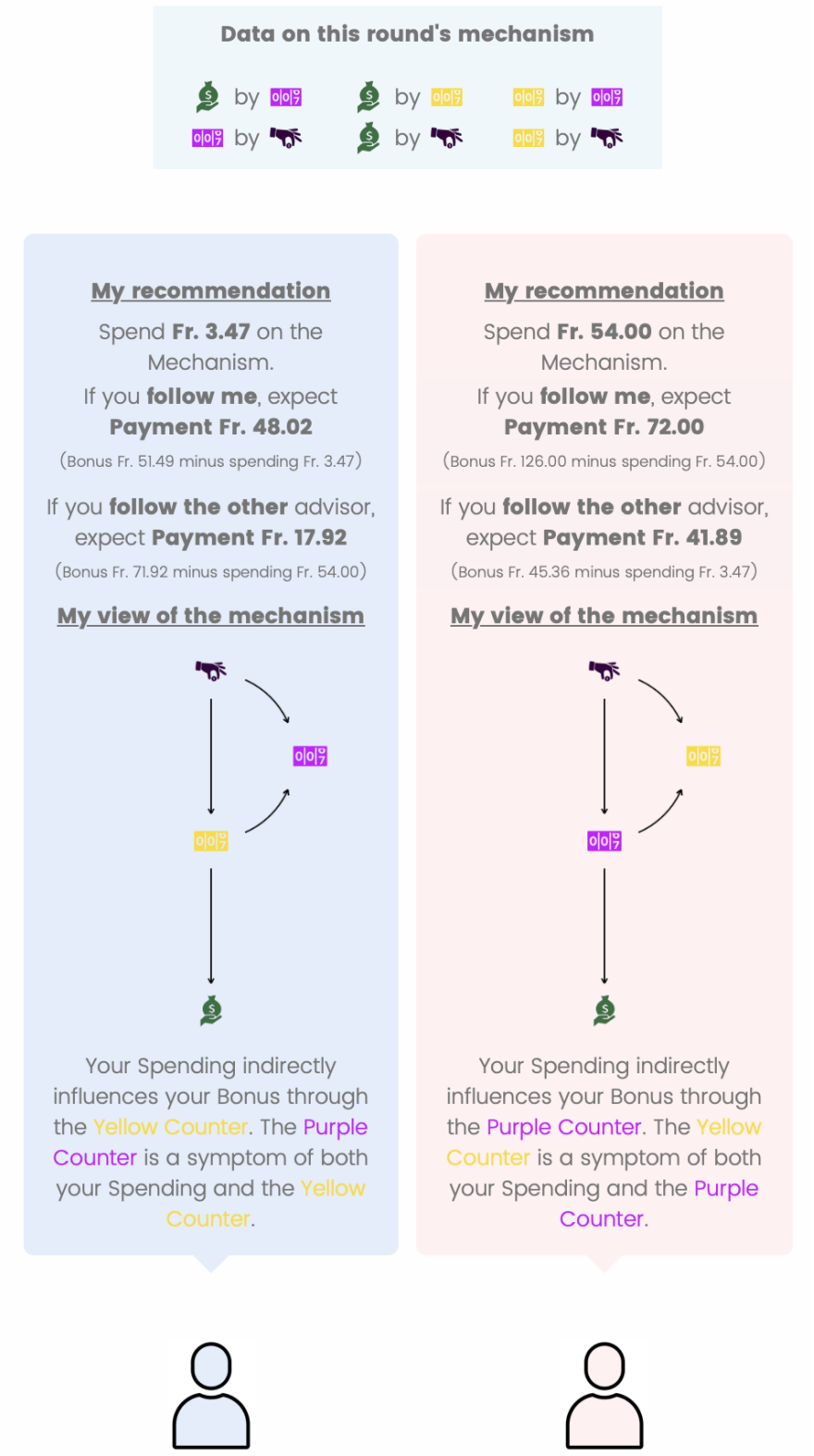
Subjects choose an investment level by deciding which of two advisors to follow, as shown in Figure 1. Each advisor has a potentially misspecified model in the form a system of regression equations that is linear in C_1, C_2, Y, and in the square root of I. The model is fitted to a large sample of data generated by the DGP (formally: the population moments implied by the DGP). We communicate models to subjects in natural language using terms such as 'X directly affects Y,' 'X influences Y indirectly through Z,' or 'X is a symptom of Y,' both in video and written format, and illustrated graphically. Each fitted model implies a *recommendation* about the payoff-maximizing investment amount and predicts the implied expected payout. It also predicts the payout the subject can expect from following the recommendation of the competing advisor. We do not explain the details of model-fitting, but we inform subjects that if an advisor's model is correct, the recommended investment is truly best for the subject, and that advisor's predictions about the payment amounts from following the own and the competing advisor's recommendation are accurate. We also tell subjects that all these numbers are meaningless if an advisor's model is incorrect.

When the subject selects an advisor, she automatically invests the amount recommended by the chosen advisor. Her payoff is determined by that amount and the DGP (with noise realizations set to zero). If the model that led to that recommendation is misspecified, the subject will thus obtain a less-than-maximal payoff.

Crucially, subjects can access data generated by the DGP to detect misspecified models. As Figure 2 illustrates, subjects can view the unconditional correlation between any pair of variables in the form of a bar chart (Panel A), and can condition each correlation on any third variable (Panel B).[7] We do not provide any hints regarding the connection between models and empirical correlations.

[7]We do not show correlations that simultaneously condition on two variables, for two reasons. First, such correlations would be needed only to test whether a diamond-shaped DAG ($I \to J \to K, I \to L \to K$, with J and L unconnected) fits the data, which we do not include in the experiment as it is not necessary to identify our decision criteria. Second, such data is rarely available public debates.

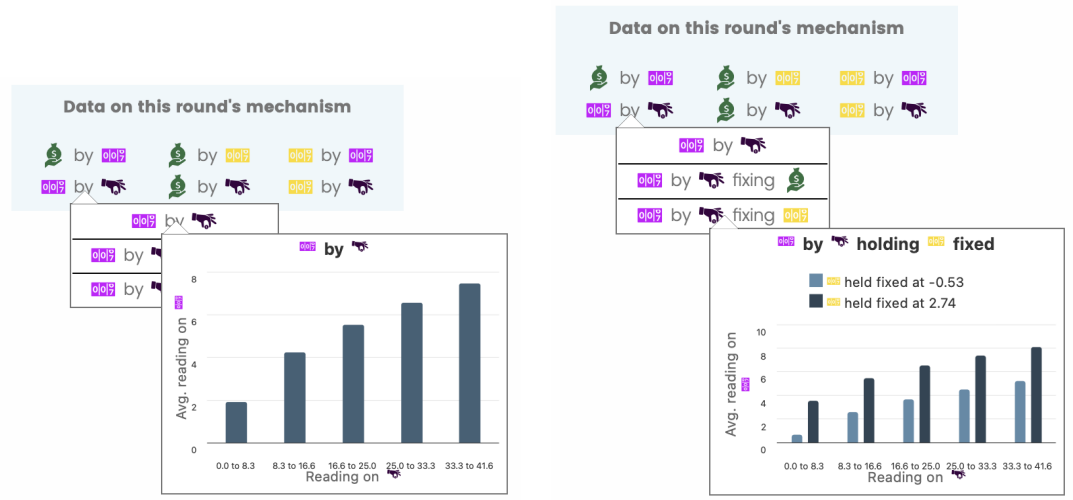
Figure 1: Decision screen



Notes: Any element in the data dashboard at the top of the screen is clickable and displays the information illustrated in Figure 2.

Figure 2: Example data charts shown to subjects

A. Unconditional correlation **B. Conditional correlation**



Notes: Subjects can retrieve each chart by clicking on the corresponding link in their 'data dashboard.' Panel A shows an example of a chart displaying an unconditional correlation. Panel B shows an example of a chart displaying a conditional correlation.

It is up to the subjects to decide which of the 18 charts to examine. It is also up to them to intuit any model's correlational implications to connect it to the data. While we illustrate models in the format of *directed acyclic graphs* (*DAG*) that represent the underlying regression specifications, we do not provide any explanation of that depiction.[8]

[8]For instance, if C_1 denotes the yellow counter, the model of the advisor on the left in Figure 1, corresponds to the three-equation system

$$\begin{aligned} Y &= \beta_Y + \beta_{C_1 Y} C_1 + \epsilon_Y \\ C_1 &= \beta_{C_1} + \beta_{IC_1}\sqrt{I} + \epsilon_{C_1} \\ C_2 &= \beta_{C_2} + \beta_{IC_2}\sqrt{I} + \beta_{C_1C_2} C_1 + \epsilon_{C_2} \end{aligned}$$

where variables β denote real-valued parameters to be estimated, and variables ϵ denote indepenent mean zero Gaussian errors whose variances may differ from each other. All our graphical illustrations of models are DAG-representations of the underlying system of regression equations. There is one equation for each variable that has one or more links pointing to it, and all variables whose links directly point into that variable (but not those that are merely indirectly connected) appear on the right-hand side of the regression equation. The interpretation is causal; a variable that appears as on the left hand side of an equation is endogenous, other variables are exogenous.

2.2 Choice criteria

We now outline the choice criteria that describe decision approaches in our setting and whose prevalence we will estimate. We first describe how correlational information can be used to rule out misspecified models and define various levels of partial inferential ability. These constitute our Fact-Based criteria. We then present the Utility-Based criteria implied by the literatures on Narrative Competition and choice under uncertainty. Because subjects may combine Fact-Based and Utility-Based criteria by first excluding misspecified models when able to and applying Utility-Based criteria otherwise, our estimation will consider all possible pairs of criteria from the two classes.

Fact-Based criteria A handful of archetypical causal models provide the key insights required for excluding misspecified models in our setting. While defined on two or three nodes, the corresponding insights also apply to the four-node DAGs we use in our experiment. Appendix A.1 provides the corresponding formal statement.

Observation 1. *Consider a system of linear Gaussian equations that is consistent with a DAG G defined over nodes $N = \{I, J, K\}$.*

(i) If $G : I \to J$, then generically $\text{cov}(I, J) \neq 0$

(ii) (a) If $G : I \to K \to J$ or $G : I \leftarrow K \to J$, then generically $\text{cov}(I, J) \neq 0$

(b) If $G : I \to K \leftarrow J$, then $\text{cov}(I, J) = 0$.

(iii) (a) If $G : I \to K \to J$ or $G : I \leftarrow K \to J$, then $\text{cov}(I, J|K) = 0$

(b) If $G : I \to K \leftarrow J$, then generically $\text{cov}(I, J|K) \neq 0$.

While causation does not necessarily imply correlation, case (i) conveys that it does so outside a knife-edge set of parameters. This fact also holds when causation is indirect (as in the *chain* $I \to K \to J$) or when a common cause affects two variables (as in the *fork* $I \leftarrow K \to J$), as stated in case (ii)(a). The independence implication in case (ii)(b) holds by definition; non-independence of I and J would require some causal path between these nodes.

To understand the conditional correlational implications in case (iii), first consider the *chain*, $I \to K \to J$. When I indirectly causes J but the mediator K is held fixed, changes in I cannot translate into changes in J, and hence $\text{cov}(I, J|K) = 0$. In the case of the *fork*, $I \leftarrow K \to J$, the common cause K is the sole reason why I and J are (unconditionally) correlated. Hence, holding K fixed eliminates that correlation, and thus $\text{cov}(I, J|K) = 0$. The implication of the *v-collider* $I \to K \leftarrow J$ becomes apparent in the example in which K is defined as $K = I + J$. Then, once we fix K, larger I must coincide with smaller J.

To see how a subject can draw inferences about model misspecification, consider the example of Figure 1. The right-hand side advisor's model contains the chain  →  →  . If the data show a correlation between  and  even if  is held fixed, the subject can infer that that model must be mistaken, following case (iii)(a) above.

Some correlational implications may be easier to intuit and understand than others. We use this variation to characterize the limits to subjects' inferential abilities:

Definition 1. ***Fact-Based criteria***

(i) Direct Links: Subjects discard a model if it posits a direct link between two nodes I and J but I and J are not correlated in the data. (Case (i) in Observation 1)

(ii) Unconditional Correlations: Subjects discard a model if any of its implied unconditional correlations (or absence thereof) are inconsistent with the data. (Cases (i) and (ii) in Observation 1)

(iii) Conditional Correlations: Subjects discard a model if any of its implied conditional correlations (or absence thereof) are inconsistent with the data. (Case (iii) in Observation 1)

(iv) All correlations: Subjects discard any model that is based on a misspecified model. (All cases in Observation 1)

Utility-Based criteria Utility-based criteria come to bear if individuals are unable or unwilling make a Fact-Based choice. These criteria depend solely on the

collection of the models' implications about payoffs and recommended investments. To define them formally, let M_1 and M_2 denote the models in the subjects' choice set. Let $M_\emptyset$ denote a third, unavailable, model in which the investment does not affect the outcome. Define a *state* of the world s that encodes whether the DGP is consistent with M_1, M_2, or $M_\emptyset$, denoted as s_1, s_2, and $s_\emptyset$, respectively. An *act* M maps states into payoffs $u_{M,s}$. The choice of model is an act.

The Narrative Equilibrium literature assumes that individuals choose the model that solves $\max_{M\in\{M_1,M_2\}} \max_{s\in\{s_1,s_2\}} u_{M,s}$—it promises the highest utility if true. Such max-max decision-making is diametrically opposed to tendency for cautious choice documented in the empirical literature of choice under ambiguity (Trautmann and Van De Kuilen, 2015). That literature suggests that a decision maker will instead select the model that promises the highest payoff if one of the non-selected models is correct. If the decision maker only considers the models available for choice, M_1 and M_2, she will solve $\max_{M\in\{M_1,M_2\}} \min_{s\in\{s_1,s_2\}} u_{M,s}$. Hence, she will first identify which model predicts the higher expected payoff from choosing its competitor, and then choose that competitor.[9] If the decision maker also accounts for the possibility that neither model is correct, and that the DGP may not feature an effect of the investment on the outcome at all, she solves $\max_{M\in\{M_1,M_2\}} \min_{s\in\{s_1,s_2,s_\emptyset\}} u_{M,s}$. In this case, the possibility that the investment will be wasted causes her to select the model that recommends the lowest investment. For completeness, we also consider the possibility that decision makers will select the model that recommends the highest investment, for instance due to the illusion of control (see, e.g., Stefan and David, 2013; Klusowski et al., 2021).[10] Overall, we thus consider four Utility-Based criteria:

Definition 2.

(i) The Best-Case Promise *criterion selects the model that promises the highest payout if correct.*

[9] When the two models are linear, this criterion is equivalent to selecting the model that promises the *lower* payoff if correct, as we formally show in Appendix A.2.

[10] Choosing the model with the highest recommended investment does not correspond to a max-max criterion. The reason is that a high investment might lead to low payoffs if it substantially exceeds the optimal investment.

(ii) *The* Worst-Case Promise *criterion selects the model that promises the highest payout if the competing model is correct.*

(iii) *The* Minimize Investment *criterion selects the model that implies the lowest investment.*

(iv) *The* Maximize Investment *criterion selects the model that implies the highest investment.*

Types We define a *type* as a pair consisting of a Fact-Based criterion and a Utility-Based criterion. A decision maker may apply a criterion from only one class. Either order of application is possible, but since each Utility-Based criterion uniquely determines choice, any subject who starts with a Utility-Based criterion necessarily ignores all Fact-Based criteria. We assume that individuals randomize uniformly across any options that remain after applying their decision criteria.

Definition 3. *A* Type *is a pair of a (possibly empty) fact-based criterion and a (possibly empty) utility-based criterion. Indeterminacies are resolved through uniform randomization.*

3 Identification and design specifics

Our experiment has two parts. The *Comprehensive Part* (Subsection 3.1) is our main focus. It identifies the frequency of types when subjects have access to correlational data. The *Utility-Focused Part* (Subsection 3.2) focuses on Utility-Based criteria by withholding the information required to apply Fact-Based criteria. It includes treatments to assess whether subjects who follow the Best-Case Promise criterion do so intentionally or by mistake. Subsection 3.3 provides details of the experimental implementation.

3.1 Comprehensive Part

The design of the Comprehensive Part follows the standard revealed-preference logic: each subject makes a choice from a sequence of menus constructed so that the overall choice distribution reveals the distribution of criteria used. Formally,

we will estimate type frequencies using a finite-mixture approach closely related to the Strategy Frequency Estimation Method of Dal Bó and Fréchette (2011, 2019).[11] Out-of-sample prediction analysis addresses concerns about overfitting.

While the approach is conceptually simple, constructing a sequence of menus that identifies the frequency of all types is challenging. Identification requires variation in observed and implied correlations as well as in recommendations and predictions. To identify the share of subjects whose choices follow the Best-Case Promise criterion, for example, we would ideally vary only which model makes the higher promise while keeping everything else constant. The fraction of subjects whose choices change in response would then indicate the prevalence of that criterion. However, we cannot vary such properties independently, since we can choose only model structures and the DGP, not the resulting identifying properties. For instance, varying the set of correlations that can inform choice requires changing a model's structure, which in turn necessarily alters its recommendations and predictions.

Our construction of a sequence of menus for identification relies on two insights. First, assessing whether a DAG fits the data requires checking only whether the predicted conditional (in)dependence relations hold, not the magnitudes of these relations. Thus, inferences based on any Fact-Based criterion are unaffected by the DGP's parameters as long as its structure remains unchanged. Once a DGP and a misspecified model are fixed, we can therefore freely choose the DGP's parameters to identify Utility-Based criteria. Second, as we show in Appendix A.2, for any pair of models we can vary independently whether the model recommending the higher investment also makes the higher or lower promise by adjusting the distribution of investments in the simulated empirical data used for model fitting. This strategy allows us to distinguish, for example, the Worst-Case Promise criterion from the Minimize Investment criterion.

[11] Unlike Dal Bó and Fréchette (2011, 2019), some types in our setting predict indifference on certain choice sets, and we estimate our model using GMM rather than MLE.

Table 1: Menus

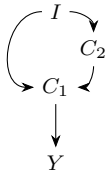
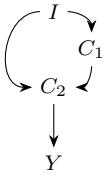
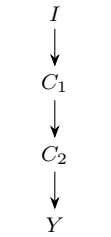
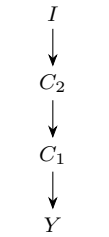
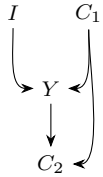
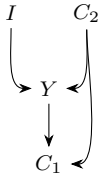
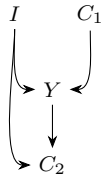
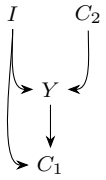
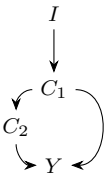
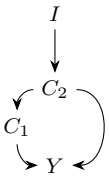
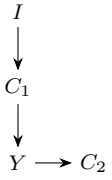
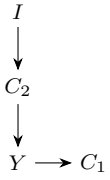
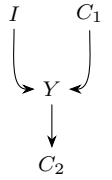
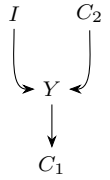
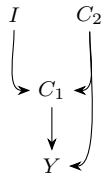
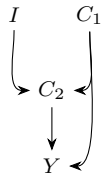
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Menu	DGP	Competitor	Description	Criterion identifies misspecified model			Model chosen by	
				Direct Links	Uncond. corr.	Cond. corr.	Best-Case Promise	Min. Invest.
M_1			Your Bonus only depends on one of the Counters. Your Action influences that Counter both directly and through the other Counter.	No	No	Yes	DGP	DGP
M_2				No	No	Yes	Comp.	DGP
M_3			Your Action indirectly influences your Bonus. It influences the first Counter which, in turn, influences the second Counter, which then influences your Bonus.	No	No	Yes	DGP	Comp.
M_4			Your Action influences your Bonus directly, as does one of the Counters. That counter is not influenced by anything. The other Counter is a symptom of both the first Counter and your bonus.	No	Yes	No	DGP	Comp.
M_5				No	Yes	No	Comp.	Comp.
M_6			Your Action influences your Bonus directly, as does one of the Counters. That counter is not influenced by anything. The other Counter is a symptom of both your Action and your Bonus.	Yes	Yes	No	DGP	Comp.
M_7				Yes	Yes	No	Comp.	Comp.
M_8			Your Action directly influences one of the Counters. That Counter influences your Bonus both directly and through the other counter.	No	No	Yes	Comp.	DGP
M_9				No	No	Yes	DGP	DGP

Table 1: Menus (continued)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Menu	DGP	Competitor	Description	Criterion identifies misspecified model			Model chosen by	
				Direct Links	Uncond. corr.	Cond. corr.	Best-Case Promise	Min. Invest.
M_{10}			Your Action indirectly influences your Bonus through one of the Counters. The second Counter is a symptom of the Bonus.	No	No	Yes	DGP	Comp.
M_{11}			Your Action influences your Bonus directly, as does one of the Counters. That counter is not influenced by anything. The other Counter is a symptom of your Bonus.	No	Yes	Yes	DGP	Comp.
M_{12}				No	Yes	Yes	Comp.	Comp.
M_{13}			Your Bonus directly depends on both Counters. Your Action influences one of these Counters, as does the second Counter. The second counter is not influenced by anything.	Yes	Yes	No	DGP	Comp.
M_{14}				Yes	Yes	No	Comp.	Comp.

Notes: I denotes the investment, referred to as Action in the videos. Column 4 shows the text spoken in the video. In the screens that correspond to Figure 1, the counters are referred to by color. In the case of M_1, for instance, the text for one of the models is "Your Bonus only depends on the red counter. Your Action influences that Counter both directly and through the blue counter.'

Three requirements lead us to impose further restrictions on the sequence of menus we construct. First, Fact-Based criteria should be applicable to distinguish between the models in each menu. Hence, the two models in each menu must stem from two different Markov-equivalence classes.[12] Second, subjects should be unable to choose based on preferences over model structures, such as favoring simpler or more complex models.[13] To achieve this, the two models in any given menu differ only in that the position of the two covariates are interchanged. Third, no two models may recommend the same investment, for otherwise subjects would have no instrumental reason to distinguish between them. Appendix A.2 shows that the set of all possible DAGs in our setting can be partitioned into 15 action-equivalence classes, with any two DAGs within the same class implying the same optimal investment and promises for any DGP. To meet the requirement, each menu contains models from two different action-equivalence classes. We also impose the technical constraints that none of our models contain isolated subsets of variables and that all have generic (non-knife-edge) parameters.[14]

After heuristically constructing a sequence of fourteen menus, we formally verify that it identifies the full vector of type probabilities, as described in Subsection 4.1. To illustrate the strength of identification achieved by this sequence, Appendix B.1 reports the distance between the choice patterns of all pairs of types.

Table 1 presents the resulting sequence $\mathcal{M} = (M_1, \ldots, M_{14})$ used to identify the distribution of types (M stands for $m$ain). Each menu in $\mathcal{M}$ includes one correct and one misspecified model. The table reports, for each menu, whether

[12] Two DAGs are Markov-equivalent if they are indistinguishable based on their conditional independence relationships; see Verma and Pearl (1990) for a characterization.

[13] An earlier version of this paper included a laboratory experiment that also measured preferences concerning the complexity or simplicity of causal structures. In that sample, 32.2% of subjects chose based on such criteria in some rounds. Overall, subjects favored more complex structures (22.9%) over simpler ones (9.3%), possibly reflecting the heuristic that models with more links seem better able to fit any data pattern. Extending the experiment to identify such *Structure-Based* criteria greatly increases the complexity of the design.

[14] A subset of variables is *isolated* if it neither influences nor is influenced by variables outside that subset. In our setting, a property of the parameter vector is *generic* if it is violated only on a subset of the parameter space with Lebesgue measure zero.

a given fact-based criterion can exclude the misspecified model and shows the choices implied by the Best-Case Promise and Minimize Investment criteria. The Worst-Case Promise and Maximize Investment criteria select the opposite model in each case. Given the restrictions imposed on the menus, we can distinguish 17 types.[15]

In addition to the 14 main menus, we include 4 menus $\mathcal{W} = (W_1, \ldots, W_4)$ (w for "$w$rong") in which both models are inconsistent with the DGP. Randomly interspersed with the main menus, these ensure that ruling out one misspecified model does not imply that the remaining model is correct. They will also provide insight into the nature of subjects' decision making; see Section 4.2. Two additional practice menus, $\mathcal{P} = (P_1, P_2)$, familiarize subjects with the decision interface and are always presented first. As preregistered, we exclude choice data from the practice menus from analysis.

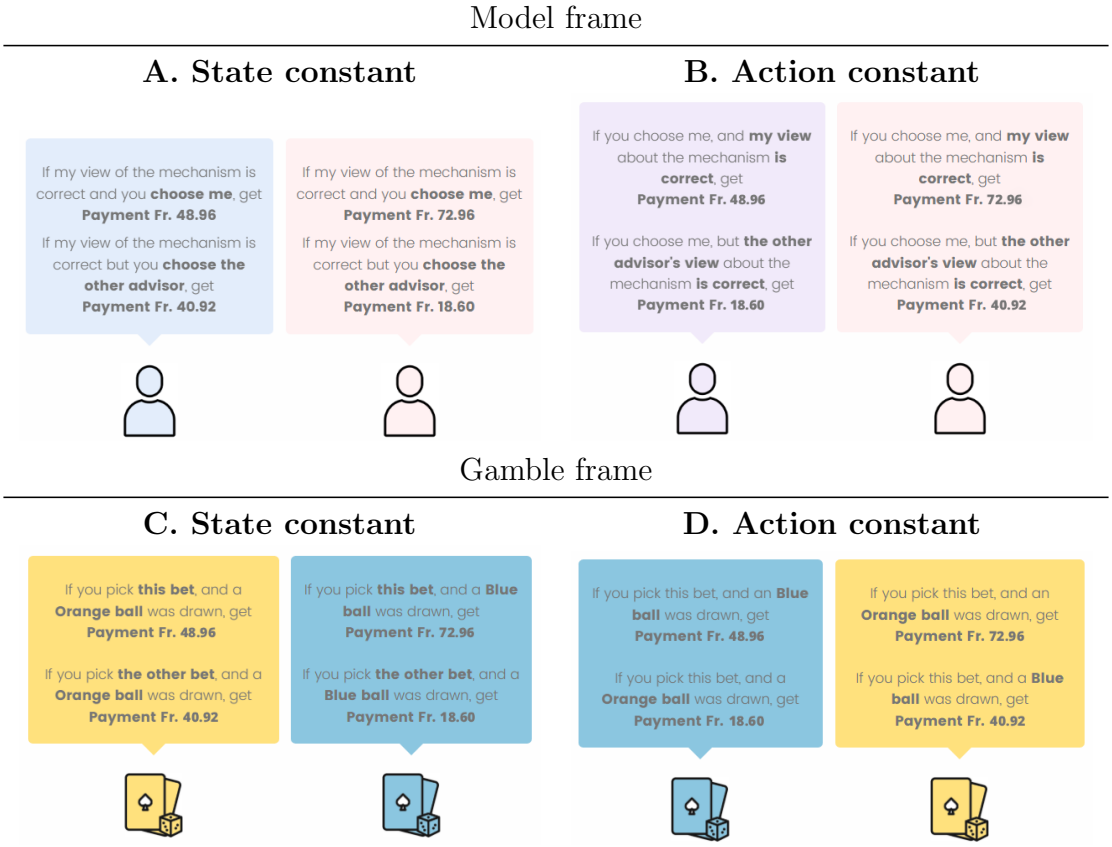
3.2 Utility-Focused Part

The Utility-Focused Part withholds the model descriptions and empirical data necessary for applying Fact-Based criteria. To avoid the need to account for the support of subjects' beliefs, which distinguishes the Worst-Case Promise and Minimize Investment criteria, we inform subjects that one model in each round is correct, and we withhold information about the recommended investment amounts. As in the Comprehensive Part, variation in the menus across multiple rounds reveals choice criteria.

If the Best-Case Promise criterion finds empirical support, the question arises why our results differ from research on decision making under uncertainty, which typically finds a preference for cautious choice. We consider two hypotheses. First, Best-Case Promise choices may occur when uncertainty arises from incomplete

[15]There are 4 Utility-Based and 4 Fact-Based criteria. Including randomization within each class increases these numbers to 5 and 5, respectively. A type who always identifies the correct model never reveals a Utility-Based criterion, implying a maximum of 21 types. Our model cannot separately identify a fully random type from a noisy population that follows nontrivial criteria, so we assume no subject randomizes throughout, leaving 20 types. Moreover, for subjects using the Maximize Investment criterion, we cannot distinguish whether they apply it alone or in combination with an imperfect Fact-Based criterion (Direct Links, Unconditional Correlations, Conditional Correlations). We attribute such choices to the type using the Maximize Investment criterion without any Fact-Based component, yielding 17 types in total.

Figure 3: Decision screen in the Utility-Focused part



knowledge about a knowable process (epistemic uncertainty), even if they do not occur in settings characterized by intrinsic randomness (aleatory uncertainty), where cautious choice is most often observed.[16] To test this hypothesis, subjects make choices in two frames. The *Model frame* mirrors Part 1 of the experiment (epistemic uncertainty; Panel A of Figure 3). The *Gamble frame* presents structurally identical choices in a balls-and-urns setting (aleatory uncertainty; Panel D). Subjects learn that either a blue or an orange ball will be drawn from an urn of unknown composition and choose between a bet on blue or and a bet on orange.

The second hypothesis is that choices consistent with the Best-Case Promise criterion reflect a failure of contingent reasoning (Niederle and Vespa, 2023). This failure may arise when model predictions are communicated in a way natural to the task but different from how uncertainty is typically presented in the litera-

[16]Prior work, if anything, finds that people are more *reluctant* to bet under epistemic than under aleatory uncertainty (Fox and Ülkümen, 2011).

ture. Using the notation of Subsection 2.2, the predictions of model i fit to data are naturally expressed as what the subject can expect, according to that model, from following its recommendation $(u(a_i^*, s_i))$ and from choosing a different action $(u(a_{-i}^*, s_i))$, as shown in Panel A of Figure 3. We refer to this as the *State-Constant* presentation. Crucially, these two values do not describe the payoff distribution the subject faces when choosing model i: if she chooses model i but the competing model is correct, her payoff is $u(a_i^*, s_{-i})$, which is a prediction of model $-i$, not of model i. Determining the possible payoffs from choosing a model thus requires contingent reasoning. Hence, subjects may misinterpret or neglect worst-case payoffs. To test this hypothesis, the *Action-Constant* presentation (Panel B) directly communicates the payoffs from choosing model i, namely $\big(u(a^i, si), u(a^i, s{-}i)\big)$. For completeness, we also include a State-Constant presentation for choices in the Gamble frame (Panel D).

A difference in choices between the State-Constant and Action-Constant presentations alone does not reveal which presentation reflects subjects' true preferences. To address this question, we vary the payout structure. In the *Spread* condition, payouts mirror those in Part 1 of the experiment: one option offers a better upside but a worse downside than the other, as necessarily implied by fitting linear models to data (see Appendix A.2). In the *Dominance* condition, one option offers both a better upside and a more favorable downside and is therefore first-order stochastically dominant. We will attribute the choice of a dominated option to a failure of contingent reasoning induced by the presentation mode.[17]

Overall, the Utility-Focused part of our experiment thus follows a $2 \times 2 \times 2$ design, as we vary the frame (Model frame vs. Gamble frame), the presentation mode (State Constant vs. Action Constant) and the payment vectors (Spread vs. Dominance). Subjects encouter each menu twice, with minor variation in payoffs, and thus make decisions in a total of 16 rounds. In one version of each menu, we flip the location of the payoff information on the screen in one but not the other speech bubble.

[17] Choosing a dominated option indicates misunderstanding of worst-case payoffs rather than mere neglect of them. Neglect would still lead subjects to select the dominant option.

3.3 Experiment implementation

All subjects first complete the Comprehensive Part, which is our main focus and cognitively more demanding than the Utility-Focused Part, which subjects complete second. The experiment then proceeds with additional elicitations. Here, we explain key design specifics. We defer details to Appendix C. Appendix E provides screenshots of the complete study interface.

In the Comprehensive Part, we ensure that subjects pay attention to the models by starting each round with a video. For each menu, the video plays a spoken version of the explanation in column 4 of Table 1 and gradually builds the graphical representation shown in columns 2 and 3.[18] The instructions emphasize that in some rounds both advisors are wrong, that the data do not affect advisors' model specifications, that some advisors' models may conflict with the correlational data generated by the DGP, that recommendations and promises result from fitting the model to the data, and that the model fitting contains no errors, yet that recommendations and promises from misspecified models are nevertheless incorrect. They also require subjects to open at least one chart chart showing unconditional correlations and one showing conditional correlations. To ensure comprehension of the instructions and the payoff structure, subjects can only continue once they pass two comprehension checks that are difficult to answer correctly by chance. The comprehension checks contain no reference to the connection between models and data.[19]

To maximize clarity, we adopt four design choices. First, the data charts omit statistical uncertainty, which could distract from our main research questions and is seldom shown in data visualizations aimed for the general public. Second, to abstract from overfitting, the charts display the true correlations implied by the DGP rather than finite-sample estimates. Third, we choose DGP parameters that yield few negative correlations, as such correlations may be harder for subjects to process, especially when chained in sequence. Fourth, we ensure that recommendations and promises differ clearly within each menu but are as similar as possible

[18] The video constructs a single DAG, with two indistinguishable grey counters indicating that the models differ only in which counter occupies which position.

[19] In each check, subjects classify each of eight statements as correct or incorrect. If they err, they are told only that a mistake occurred, not which or how many statements are wrong, and must revisit the instructions until passing.

across menus. In the High Stakes condition, one model generally promises around Fr. 72 from following it and around Fr. 18 from following the competitor, while the other promises around Fr. 48 from following it and around Fr. 42 from following the competitor. Amounts are reduced by two-thirds in the Low Stakes condition.[20]

The Utility-Focused Part begins with instructions and is block-randomized: some subjects first make all decisions in the Model frame, others first make all decisions in the Gamble frame. Within each frame, subjects first complete decisions either in the State-Constant or in the Action-Constant presentation mode, and then switch to the respective other presentation mode. Each of these four blocks begins with a screen explaining the payoff format. Subjects must correctly indicate how much they would earn from choosing the option on the left if the option on the right turned out to be ex-post payoff-maximizing. The possible payouts mirror those in the Comprehensive Part. In the *Spread* condition, the payoff vectors are (72, 18) and (48, 42). In the *Dominance* condition, they are (72, 42) and (48, 18). To avoid repetition across rounds, we uniformly randomly vary payment amounts within Fr. 1.20 of these values. All amounts are reduced by two-thirds in the Low Stakes condition.

Toward the end of the study, we elicit several individual characteristics to relate to the use of decision criteria. These include each subject's field of study (classified as STEM, economics and business, or other) and measures of familiarity with concepts in probabilistic causal inference: completing the aphorism "correlation does not…," writing out the name of the mathematical object $P(A|B)$ in words, spelling out the acronym "DAG," and reporting whether they have taken a class on causal statistical inference.[21]

We also elicit risk preferences following Eckel and Grossman (2008) and ambiguity preferences following Dimmock et al. (2015), both incentivized. To reduce noise and allow OR-IV estimation (Gillen et al., 2019), subjects complete two versions of each elicitation with slightly different parameters. We also administer an extended Cognitive Reflection Test (Frederick, 2005; Thomson and Oppenheimer,

[20] Appendix Table C.4 lists the precise amounts for each menu. In some cases, it was not possible to choose DGP parameters to produce precisely these amounts as model implications.

[21] We score the first three items by whether responses include the strings "caus," "conditional" or "given," and "acyc," respectively.

2016) and measure belief in pseudoscience (Torres et al., 2020). Subjects report their gender and the Swiss political party closest to their views, which we locate on the political spectrum using Jolly et al. (2022).[22] Finally, subjects describe in their own words how they typically made decisions in the main rounds of the experiment.

Subjects learn that a single decision from the entire study will be randomly selected for payment. To eliminate the influence of standard risk preferences, subjects are paid the expected value of their chosen option according to the DGP. The model's stochasticity serves solely to generate variation for data fitting.

4 Analysis

We conducted the experiment with 414 subjects in June and July 2025 at the Laboratory for Experimental and Behavioral Economics at the University of Zurich. The mean payment was Fr. 47.63 (USD 60.00), with averages of Fr. 31.96 and Fr. 63.78 in the Low and High Stakes conditions, respectively.[23] Appendix D.1 provides summary statistics of our sample's composition. Subjects were required to spend at least 75 minutes in the lab. Some took up to 2 hours, with a median completion time of around 80 minutes. The median time spent on a main round was 50 seconds.[24]

We organize the analysis into five subsections. The first three concern the Comprehensive Part. Subsection 4.1 documents subjects' remarkable ability to discard misspecified models and presents the main estimates of all decision criteria. Subsection 4.2 focuses on menus in which all available models are misspecified to show that subjects draw inferences by detecting qualitative inconsistencies between model implications and empirical data rather than through quantitative reasoning. Subsection 4.3 uses out-of-sample prediction analysis to demonstrate the robustness of our results. It also provides parsimonious models whose pre-

[22] The survey lacks a score for the Swiss Communist Party ("Partei der Arbeit"); we assign it a value of 0, the leftmost position.

[23] We preregistered a target sample size of 400 subjects, https://www.socialscienceregistry.org/trials/16173.

[24] The median time spent on a practice round was 119 seconds. Appendix D.2 examines order effects. Minor order effects in terms of time spent per round appear to reflect learning rather than fatigue, as subjects viewed data charts at a constant rate throughout the experiment.

dictive accuracy rivals that of our unrestricted model. Subsection 4.4 studies the Best-Case Promise criterion by analyzing choices from the Utility-Focused Part of the experiment. Finally, Subsection 4.5 relates decision-making to subjects' educational background, political preferences, and demographic and psychological traits.

4.1 Comprehensive Part

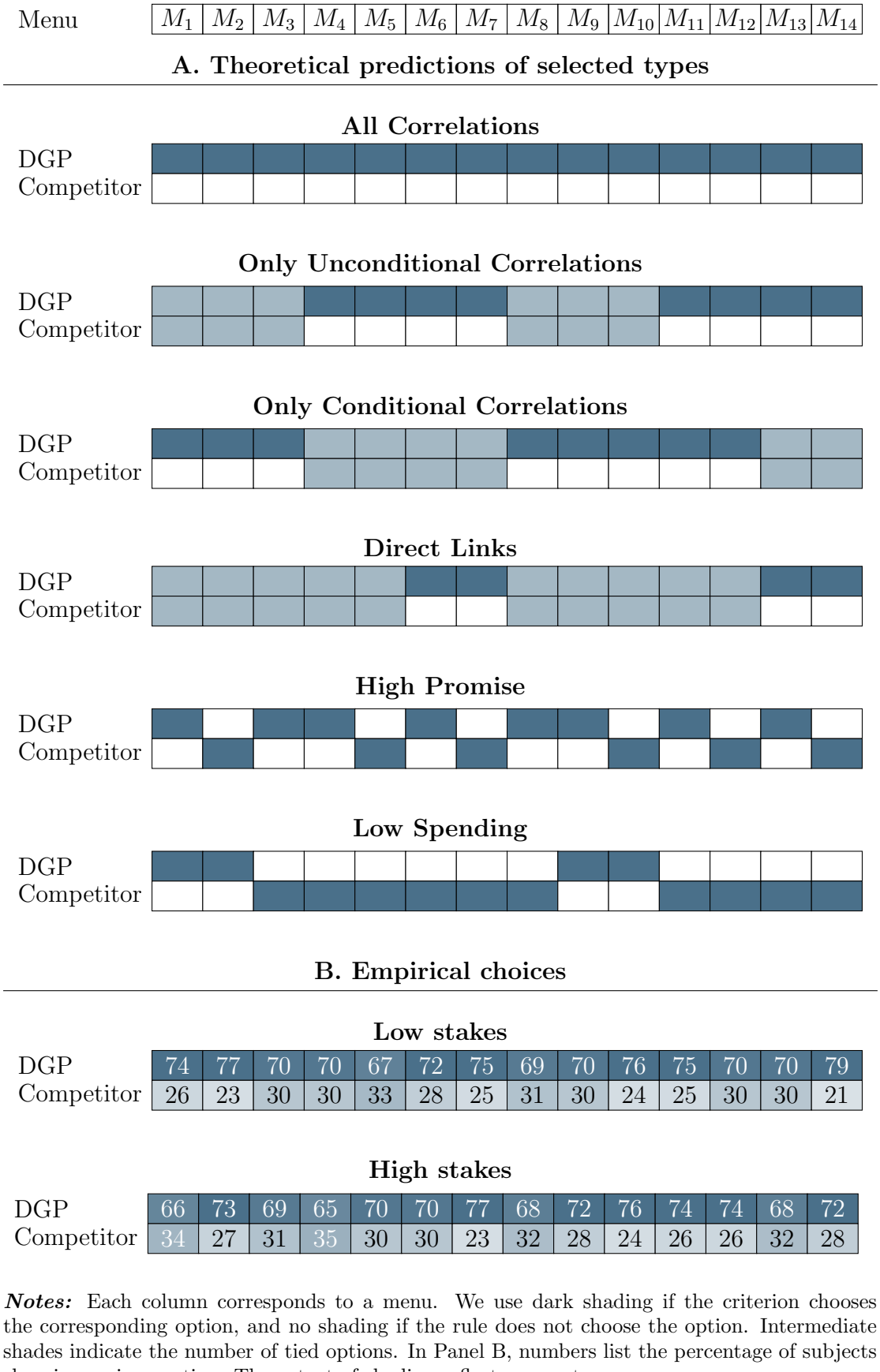
Our main evidence consists of the incentivized choices across the 14 menus in set $\mathcal{M}$. To help interpret these choices, Panel A of Figure 4 shows the predicted choice distributions for selected types across the menus. A subject able to link each model's correlational implications to the data will always select the correct model. In contrast, a subject who processes only unconditional correlations will select the correct model in menus M_4 to M_7 and M_{11} to M_{14}. For the remaining menus, light blue shading indicates uniform randomization, reflecting the randomization assumption made in Section 2.2.

Panel B shows the aggregate empirical choice distributions for the Low and High Stakes conditions. Three patterns emerge. First, subjects often choose the correct model. The aggregate data most closely match the type that consistently selects the correct model. In fact, in each round, at least two thirds of subjects choose the correct model, and the average subject does so in 10 of the 14 rounds—well above the random-choice benchmark of 7. Second, choices of misspecified models are also common (about one fifth to one third of cases). These choices vary systematically across menus but do not resemble any single type's pattern, indicating heterogeneity across subjects. Third, choice distributions are similar across stakes conditions: higher stakes do not lead subjects to select the correct model more often, if at all. Hence, failures to identify the correct model reflect limited ability rather than lack of motivation.[25]

Mixture model To draw detailed inferences about the use of each decision criterion, we fit a finite mixture model that yields an estimate $(\hat{t}_1, \ldots, \hat{t}_n)$ of the

[25]Pilot experiments on Prolific also showed no stake effects. However, more than 80% of those subjects refused to view even a single data chart, apparently attempting to rush through the study.

Figure 4: Fingerprints of selected choice criteria



Notes: Each column corresponds to a menu. We use dark shading if the criterion chooses the corresponding option, and no shading if the rule does not choose the option. Intermediate shades indicate the number of tied options. In Panel B, numbers list the percentage of subjects choosing a given option. The extent of shading reflects percentages.

frequency t_i of each type i in our data. Conceptually, summing the theoretically predicted choice distributions in Panel A of Figure 4, weighted by a candidate type distribution $\tilde{t} = (\tilde{t}_1, \ldots, \tilde{t}_n)$, gives the aggregate choice distribution we would expect to observe. The estimation procedure searches for the distribution $\tilde{t}$ that minimizes the distance between predicted and observed aggregate choice distributions. To allow for stochastic choice, we assume that in each round a subject either selects a type-consistent option with probability $(1 - q)$ or randomizes uniformly and independently across all options with probability q, following Costa-Gomes and Crawford (2006); Ambuehl and Bernheim (forthcoming).[26]

We define the distance between the predicted and observed choice distributions as the weighted sum of all first and second moments. The first moments are the differences between predicted and observed probabilities of choosing the correct model for each menu m. The second moments are the corresponding differences conditional on choices in another menu m', for all menu pairs (m, m').[27] We estimate the model using the generalized method of moments (GMM), applying the standard two-stage feasible GMM procedure to obtain optimal moment weights. Appendix B.2 provides details.

Our mixture model has three attractive features. First, it guarantees that the estimated type distribution converges to the true population distribution as the number of subjects increases. Individual-level approaches such as Bayesian classifiers offer no such guarantee when the number of decisions per subject is fixed. They may therefore perform poorly when choices are noisy, as our simulations confirm. Second, holding the stochastic choice probability q fixed, the predicted aggregate choice distribution is a linear function of the type frequencies. This linearity allows analytical proof of identification: a sequence of menus identifies

[26] We designed the experiment to make this assumption plausible. Subjects saw menus in individually randomized order, ensuring that any inattention is evenly distributed across menus. The screen position of models was also randomized, so tendencies such as always choosing the right-hand option appear as uniform randomization across menus.

[27] Second moments are needed for identification. Suppose there are two menus, each with two options (A, B), and three types: type 1 always chooses A, type 2 always chooses B, and type 3 randomizes. If A is chosen 50% of the time in both menus, first moments alone cannot distinguish between a population of 50% type 1 and 50% type 2, a population of only type 3, or a mixture of the two. Including second moments resolves this, as only type 3 produces the pattern of choosing A in one menu and B in the other.

the vector of type weights if this linear function has full rank.[28] Third, because of linearity, the GMM objective function is quadratic for any fixed noise probability. Numerical optimization is therefore rapid and will not get stuck in local optima, even for large type sets.[29] The model has two limitations. First, it assumes that the mean noise probability is identical across types. Second, the stochastic choice probability and the share of subjects who randomize uniformly throughout are not separately identifiable. We interpret the data under the assumption that the latter share is zero.

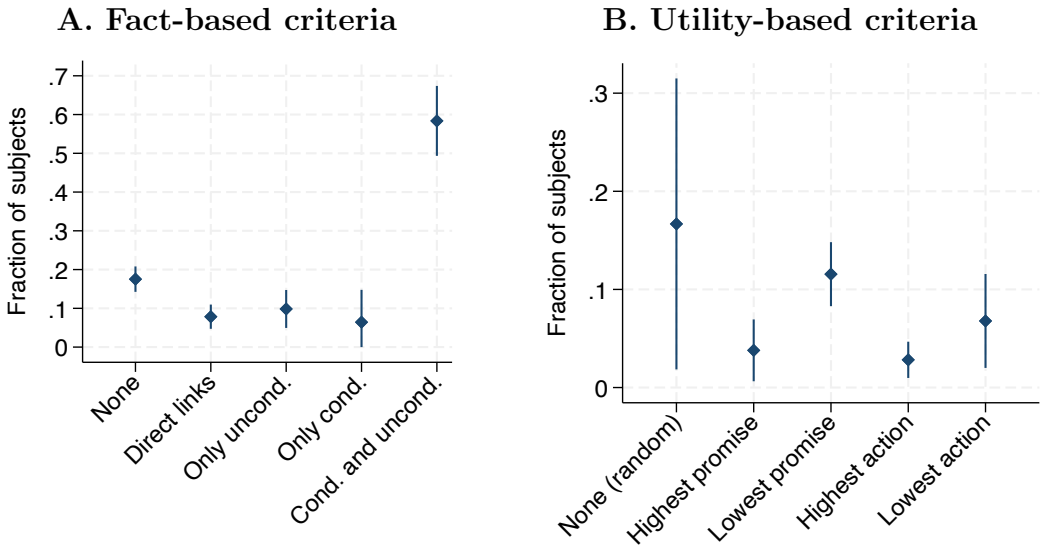
Model estimates Our estimation results confirm the visual impression from Figure 4: in both the Low and High Stakes conditions, the most common type draws correct inferences from both conditional and unconditional correlations. This type accounts for 65.4% (s.e. 4.9%) and 56.9% (s.e. 4.8%) of subjects in the Low and High Stakes conditions, respectively. These shares do not imply that more than half the subjects identify the correct model in every round. The estimated noise parameters are 24.2% (s.e. 2.0%) and 21.4% (s.e. 2.0%) in the Low and High Stakes conditions, respectively. Thus, even subjects who generally identify the correct advisor randomize in roughly one out of five rounds and thus select the wrong model in about one out of ten, with such errors distributed randomly across menus and options. Each remaining type accounts for less than 10% of subjects, though many receive nontrivial weight. Appendix D.3 reports the complete estimated type vector.

To summarize the remaining subjects, recall that each type combines a Fact-Based and a Utility-Based criterion. We estimate the prevalence of each Fact-Based criterion by summing over all types that apply it. Likewise, marginalizing over Fact-Based criteria yields the distribution of Utility-Based criteria. Aggregating across types also helps average out noise that could affect estimates when many types are fitted to limited data.

[28] We prove identification with the noise probability held fixed and verify through simulations that the estimator correctly recovers the noise probability in synthetic data.

[29] Because these properties hold only when the noise probability is fixed, we define a grid of starting values for it, run the minimization for each, and select the global optimum.

Figure 5: Distribution of decision criteria



Notes: Pooled across stakes, and training and test sets. Whiskers show 95%-confidence intervals, truncated at 0. *Panel A: Panel B:* Estimates of advice-based criteria in Experiment 1 are shown conditional on not using the Conditional Correlations criterion.

Panel A of Figure 5 plots the distribution of Fact-Based criteria using all 14 rounds and pooling across the stakes conditions.[30] Next to the majority of subjects who consistently identify the correct model, 24.1% make partly fact-based decisions. While similar numbers of subjects use each of the partial Fact-Based criteria, two of these three criteria only require an understanding of unconditional but not conditional correlations. The remaining 17.5% (s.e. 1.7%) use no Fact-Based criterion at all.

Panel B displays the distribution of Utility-Based criteria, focusing on subjects who do not consistently select the correct model. Support for the Best-Case Promise criterion is minimal (3.8%, s.e. 1.6%). Instead, subjects who deploy Utility-Based criteria optimize worst-case outcomes (18.3%, s.e. 3.1%), either through the Worst-Case Promise criterion (11.6%, s.e. 1.7%) or through the Minimal Investment criterion (6.8%, s.e. 2.4%). Nonetheless, among subjects who fail to identify the correct model through Fact-Based criteria in some rounds, the

[30] Appendix D.4 shows these distributions estimated only using the training set and split by stakes condition, as preregistered. The qualitative results are unchanged, although the use of a much smaller sample greatly increases the standard errors of the estimates.

modal choice (17.5%, s.e. 1.7%) is uniform randomization, consistent with the principle of insufficient reason.

Decision strategies What strategies help subjects identify the correct model? We argue that subjects rely on qualitative inference, looking for inconsistencies between model implications and the data. A competing hypothesis is that subjects rely on quantitative inference by visually estimating the return to investment from the chart showing the relation between investment and bonus. While a few subjects report doing so when asked to describe their decision-making in words, viewing data show this behavior is rare. Subjects view the investment/bonus-chart exclusively in only 2.1% of cases. Moreover, they view it at all in just 26.2% of cases, compared with 42.4% and 41.7% for the charts relating investment to each covariate, respectively. For subjects seeking qualitative inconsistencies between model implications and data, the latter charts are useful in many rounds, whereas the investment/bonus-chart never serves that purpose.

The fact that subjects frequently choose the correct model despite receiving no assistance suggests that the qualitative, model-based inference in our experiment comes naturally to many people, unlike the purely quantitative reasoning required in standard balls-and-urns belief-updating tasks.

How do subjects draw these inferences? Do they reason from data to models, viewing charts until a pattern stands out to them, or from models to data, intuiting the models' correlational implications and checking the data for those that distinguish the models? Several observations support the latter hypothesis. Of the 18 charts available per menu, the median subject views only four, consistent with the idea that they know what to look for. Of the 12 charts showing conditional correlations, the average subject views 2.84, of which only 0.68 do not distinguish the models. Viewing non-distinguishing charts can still be useful, as it helps detect when both causal models in a round are misspecified. Targeting is less precise for unconditional correlations, likely due to individual heterogeneity, but still consistent with purposeful selection: of the 6 charts showing unconditional correlations,

subjects view on average 1.85, roughly half of which (0.87) distinguish the models. This far exceeds what would be expected if chart selection were random.[31]

Subjects' general sophistication A potential criticism of our interpretation is that our participants' performance reflects an unusual level of sophistication and therefore may not generalize to other subject pools. Our data cast doubt on this view. Although a substantial share of subjects major in STEM (64%), their statistical sophistication is limited. Only 54% can complete the aphorism "Correlation does not...," 22.5% report having taken a class on statistical causal inference, one quarter can name the expression '$P(A|B)$,' and 8% can spell out "DAG." A comparison of our subjects' mean CRT score of 71.4% to that of other subject pools suggests a similar conclusion. Students at TU and LMU Munich achieve 75%, (Coutts et al., 2025), and those at Mannheim University and Osaka University score 68% and 83%, respectively (Glaser et al., 2019; Hanaki et al., 2021). While these scores are higher than those found in representative U.S. samples (58.3%, Caplin et al., 2023), they fall short of scores observed among professionals in the financial (Thoma et al., 2015; Angrisani et al., 2022; Weitzel et al., 2020) and oil-producing (Welsh and Begg, 2017) industries.[32]

Overall, we find that subjects display a remarkable ability to intuit models' correlational implications and verify them against empirical data to rule out misspecified models. Utility-Based criteria that receive empirical support maximize worst-case rather than best-case payoffs.[33] Support for the central assumption of the Narrative Competition literature, the Best-Case Promise criterion, is minimal. These conclusions remain robust to a threefold increase in stakes.

[31] In an average round, 86% of subjects view at least one data chart. Overall, subjects view at least one chart that distinguishes the models in 60.3% of rounds, consistent with our estimated Fact-Based criteria in Figure 5, given the estimated random-choice probability of 24.9%.

[32] Different studies use different extensions of the CRT test. While we use Thomson and Oppenheimer (2016), the studies cited above predominantly use Toplak et al. (2014), which limits comparability. (Hanaki et al. (2021) also includes questions from Finucane and Gullion (2010)). Performance on the three original CRT questions (Frederick, 2005) is available for three of these studies. Subjects in Glaser et al. (2019), Thoma et al. (2015), and Welsh and Begg (2017) answer 68%, 91%, and 81% correctly, respectively. Our subjects' score of 77.3% on those three questions falls well within this range.

[33] When asked to describe their decision process, many subjects report trying to infer the correct model from the data and defaulting to payoff safety when uncertain.

4.2 Choices when both advisors are wrong

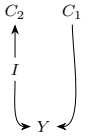
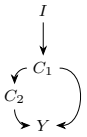
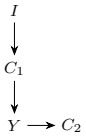
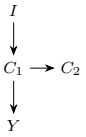
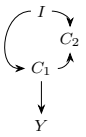
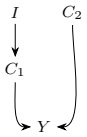
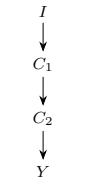
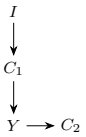
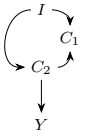
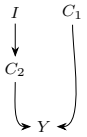
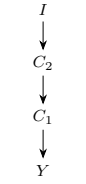
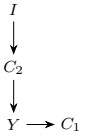
We next examine choices on menus in $\mathcal{W}$, where both models are inconsistent with the data.[34] A first finding is that the average subject does not merely select the remaining option after ruling out one misspecified model. Instead, subjects also evaluate whether the remaining models fit the data: when neither model is correct, the average number of charts viewed rises from 4.7 to 5.2 ($p < 0.01$), and average time per round increases from 50 to 68 seconds.

More importantly, choices from menus in $\mathcal{W}$ provide further evidence that subjects reason qualitatively rather than quantitatively, as follows. Once a subject determines that both available models are misspecified, several decision approaches are possible. First, she may continue reasoning qualitatively and select the model with fewer inconsistencies between the data and its correlational implications. As shown in Panel B of Table 2, this criterion applies to all menus in $\mathcal{W}$ except W_1 and requires understanding conditional correlational implications. Second, she may reason quantitatively by estimating the strength of the relationship between investment and bonus from the corresponding chart and choosing the model whose recommended investment is closer to the optimal level implied by that chart, as listed in Panel C. Third, she may revert to a Utility-Based criterion when the available Fact-Based criteria of Definition 1 do not yield a clear choice.

Panel D displays subjects' choices. In menus where one model has fewer correlational inconsistencies with the DGP, about 60 to 70% of subjects choose that model. When both models have the same number of inconsistencies (menu W_1), choices are evenly split. This pattern indicates that minimizing the number of inconsistencies is subjects' preferred criterion. In contrast, quantitative reasoning predicts that subjects predominantly choose model 2 in menus W_1 and W_2, yet they do so in only 40 to 50% of cases.

[34] The analysis in this subsection is exploratory and was not preregistered.

Table 2: Menus with two misspecified models

	W_1	W_2	W_3	W_4
A. DGP and model structures				
DGP				
Model 1				
Model 2				
Description	Your Action indirectly influences your Bonus through one of the Counters. The second Counter is a symptom of both your Action and the first Counter.	Your Action indirectly influences your Bonus through one of the counters. The other counter is not influenced by anything, but it has an additional, direct influence on your Bonus.	Your Action indirectly influences your Bonus. It influences the first Counter which, in turn, influences the second Counter, which then influences your Bonus.	Your Action indirectly influences your Bonus through one of the Counters. The second Counter is a symptom of the Bonus.
B. Inconsistencies with DGP				
Unconditional correlations				
Model 1	4	4	0	0
Model 2	4	4	0	0
Conditional correlations				
Model 1	10	2	8	6
Model 2	10	10	12	14
C. Implied optimal investment				
DGP	12.40	78.16	28.88	18.07
Model 1	0.00	8.03	24.68	18.07
Model 2	10.04	36.08	3.26	1.17
D. Choice frequencies				
Low stakes				
Model 1	51	61	69	68
Model 2	49	39	31	32
High stakes				
Model 1	50	57	62	68
Model 2	50	43	38	32

Notes: Panel A shows the structure of the DGP, and the two available models in each menu in set $\mathcal{W}$. Panel B shows the number of inconsistencies between each model and the DGP. Panel C shows the investment amounts implied as optimal by each model when fit to the DGP and by the DGP itself in the Low Stakes condition. Panel D shows empirical choice frequencies in percent.

To formally estimate the prevalence of the decision approaches, we fit a version of our mixture model on choices from the menus in $\mathcal{W}$. We exclude all Fact-Based criteria from Definition 1 since $\mathcal{W}$ is constructed such that these criteria do not apply. Instead, we include a type that minimizes the number of inconsistencies between the chosen model and the DGP (randomizing when the numbers are equal),[35] and a type that selects the model whose implied optimal investment is closest to the true value. The latter type's behavior coincides with the Maximize Investment criterion. Given its negligible incidence in Section 4.1, we set the fraction of subjects following that criterion to zero and attribute the corresponding choices to quantitative fact-based reasoning. We also include all types that apply one of the Utility-Based criteria (except the Maximize Investment criterion) from Definition 2 without combining it with any other criterion.

Table 3: Distribution of types when both models are inconsistent with the DGP

	(1)	(2)	(3)
Stakes	Low	High	p-value
Utility-Based Criteria			
Best-Case Promise	0.059	0.121	0.202
	(0.034)	(0.035)	
Worst-Case Promise	0.086	0.222	0.005
	(0.034)	(0.035)	
Minimize Investment	0.108	0.067	0.359
	(0.031)	(0.033)	
Data-Based Criteria			
Closest recommendation	0.134	0.138	0.941
	(0.037)	(0.038)	
Fewest Inconsistencies	0.614	0.453	0.000
	(0.028)	(0.035)	
Noise	0.450	0.418	0.682
	(0.055)	(0.054)	
Subjects	210	204	

Notes: Column 3 shows the p-values of a z-test of the null hypothesis that a criterion is equally common across the Low and High Stakes conditions.

[35] We assume randomization because subjects rarely combine Utility-Based and Data-Based criteria in Section 4.1.

Table 3 reports the results. Consistent with the visual evidence from Table 2, subjects who reason qualitatively by minimizing the number of inconsistencies are far more prevalent than those who reason quantitatively by selecting the model whose recommended investment is closest to the truth.

4.3 Out-of-sample predictions and parsimonious models

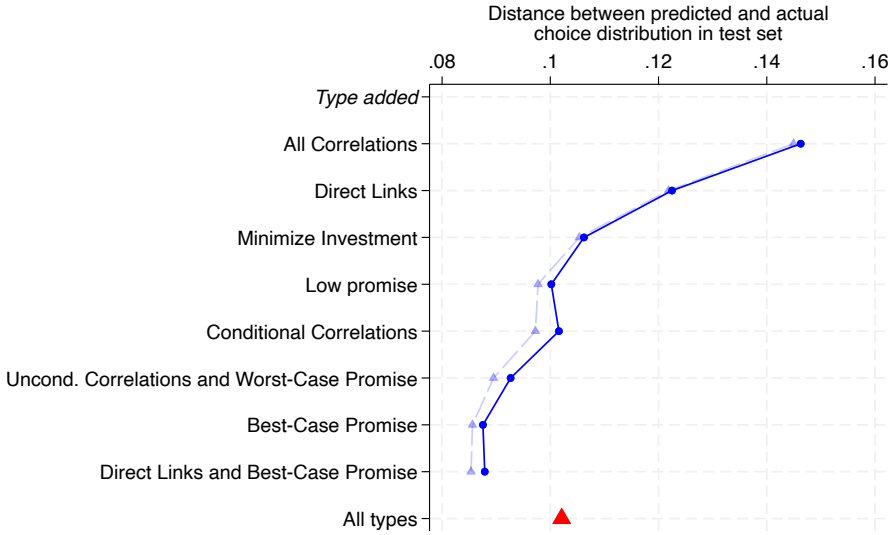
We now perform out-of-sample predictions to test whether our main classification reflects stable behavioral tendencies rather than overfitting. Relatedly, we examine whether more parsimonious models can predict choices similarly well as our unrestricted, 17-types model.

To answer our first question, we define the training set $\mathcal{M}^{train} = \{M_1, \ldots, M_7\}$ and the test set $\mathcal{M}^{test} = \{M_8, \ldots, M_{14}\}$, as preregistered. We estimate the model on the training set and then compute the Euclidean distance between the predicted and observed joint choice distributions across all menus in the test set. Importantly, each model appears only in either the training or the test set, so out-of-sample predictive power reflects general choice procedures rather than simple repetition of choices for identical models. The training set allows identification of all 17 types, while the test set identifies a 16-dimensional subset.We pool the Low and High Stakes conditions.

We find that the distance between the predicted and empirical distributions on the test set is 0.092. For comparison, the distance between the uniform and empirical distributions is 0.844. Thus, the full model reduces this distance to about one-tenth (0.108). This strong out-of-sample predictive performance indicates that the model captures stable behavioral tendencies.

To address the second question of whether a more parsimonious model can match the full model's predictive power, we evaluate all possible subsets of the 17 types. We repeat the out-of-sample prediction procedure for each subset. Because random variation in the test set could inflate the apparent performance of specific type combinations, overfitting is a concern. To mitigate it, we apply a bootstrap approach. Instead of using the original sample, we run the procedure on each of 100 bootstrap resamples of subjects. For each $k \in 1, \ldots, 17$ and each bootstrap sample b, we identify the subset S_k^b among all k-element subsets that yields the

Figure 6: Out-of-sample predictive power as function of the number of types



Notes: This chart plots the Euclidian distance between the predicted and observed marginal distributions of choices on the test set, normalized by the distance between the uniform and observed distributions. The solid line averages only across bootstrap samples in which the listed typeset predicts best, that is samples b with $S_k^b = S_k$. The dashed line averages the normalized distance across all bootstrap draws, including those for which the most predictive type set differs from S_k.

highest predictive accuracy for that bootstrap sample. We then aggregate across bootstrap samples by selecting the k-element subset S_k that performs best in the largest number of bootstrap draws.

We find that the sets of best-predicting types are nested when we include 8 or fewer types. For each k, the best-predicting k-element set S_k consists of the best-predicting $(k-1)$-element set S_{k-1} plus one additional type. Figure 6 shows the sequence of added types for each k up to 8 and displays the distance between the predicted and empirical distributions, normalized by the distance between the uniform and empirical distributions on the test set.

We find that the largest gain in out-of-sample predictive power comes from including a single type in the type set. As our analysis in section 4.1 already suggests, this is the type that consistently selects the correct model. It alone reduces the distance between predicted and actual distributions by more than 85%. The best two-type combination adds the type that exclusively follows the Direct Links criterion. Utility-Based criteria appear once we include three or more types. The first additions capture cautious choice, with the Minimize Investment

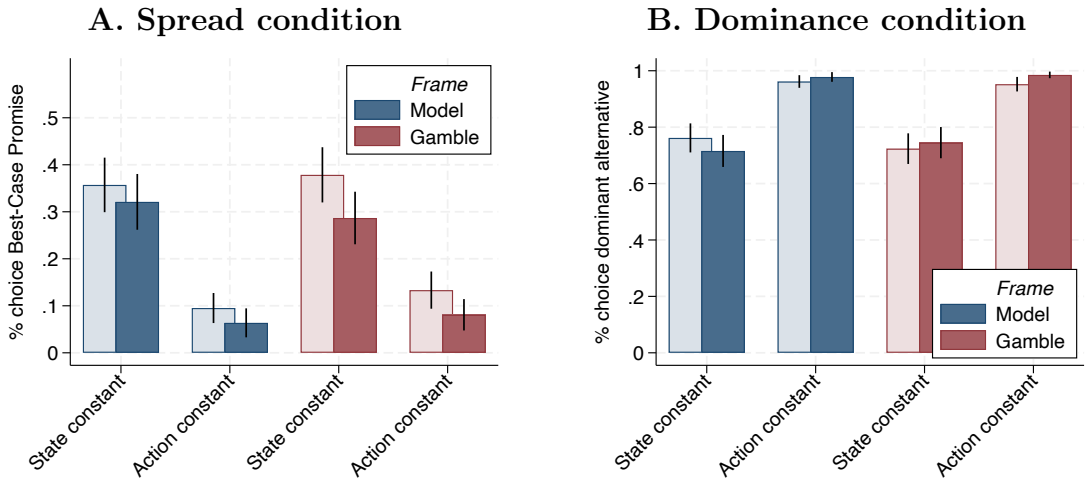
criterion in the three-element set and the Worst-Case Promise criterion in the four-element set. The best three-element set rivals the predictive power of the full model, and the best four-element set exceeds it. Predictive accuracy continues to improve up to eight types and then declines.[36] This decline is expected, as larger sets increase the risk of overfitting, which produces more variable and thus weaker out-of-sample predictions.

We conclude that a three- or four-type model consisting of a fully accurate type, a type that learns only from correlations corresponding to direct links in the models, and a type that maximizes worst-case outcomes without consulting the data captures most of the predictable variation in the data.

4.4 Utility-Focused Part

We now examine how individuals choose absent access to the information required to apply Fact-Based criteria. This analysis provides insight into the Best-Case Promise criterion and thus speaks to the Narrative Competition literature.

Figure 7: Utility-based choices



Notes: Light bars represent the Low Stakes condition, dark bars the High Stakes condition. Whiskers represent 95% confidence intervals with standard errors clustered by subject.

[36] Optimal model complexity depends on sample size, see, for instance, Montiel Olea et al. (2022).

Figure 7 summarizes subjects' choices in the Utility-Focused Part. Panel A shows how often subjects choose the Best-Case Promise alternative in the Spread condition (light and dark bars indicate the Low and High Stakes conditions, respectively). When information is presented in the format natural to models fit to data—the State-Constant presentation—about one third of subjects choose the Best-Case Promise alternative, providing some empirical support for the Narrative Competition literature.[37] However, Best-Case Promise choices fall below ten percent when payoffs are presented in the format typical of experiments on choice under uncertainty—the Action-Constant presentation.

Which presentation reflects subjects' actual preferences? Panel B addresses this question by showing the frequency of choosing the dominant alternative in the Dominance condition. Subjects almost always choose the dominant option in the Action-Constant presentation, but about one quarter fail to do so in the State-Constant presentation. Hence, choices consistent with the Best-Case Promise criterion in the Spread condition largely reflect mistakes.

While the presentation mode greatly affects choices, a change from the Model Frame to the Gamble frame has no discernible effect.[38] A threefold increase in stakes causes a minor decrease in Best-Case Promise choices. Assuming higher stakes increase effort and reduce confusion, this finding supports our interpretation that Best-Case Promise choices do not reflect preferences.

We conclude that the key assumption of the Narrative Competition literature describes a sizable minority of subjects when model implications are presented in the way natural to the setting, but these choices reflect a cognitive limitation rather than preferences.

[37] This share is sizable relative to findings in the literature on cautious choice and in the separate literature reviewed by Engelmann et al. (forthcoming), which rarely observes behavior consistent with anticipatory utility in laboratory settings.

[38] Appendix D.5 shows that all comparisons across presentation modes are highly statistically significant, that all comparisons across the Model and Gamble frames are insignificant at the 5% level, and that stake-effects are significant in some conditions. It also complements these aggregate statistics with individual-level analysis to show that the aggregates do not mask individual-level heterogeneity.

Table 4: Effects of subjects' background

	(1)	(2)	(3)	(4)	(5)
	Part 1	Part 2			
Dependent variable	Chooses correct model	Chooses maxmax		Chooses dominated	
Treatment		Spread		Dominance	
Presentation mode					
State constant	✓	✓		✓	
Action constant			✓		✓
Preference measures (OR-IV)					
Risk aversion perc. rank (0 to 1)	-0.153**	0.004	-0.230***	0.176	-0.072*
	(0.057)	(0.106)	(0.061)	(0.098)	(0.034)
Ambiguity aversion perc. rank (0 to 1)	0.078	-0.110	-0.118*	-0.034	-0.003
	(0.045)	(0.087)	(0.052)	(0.081)	(0.024)
Demographics					
Female	0.030	-0.008	0.011	0.022	-0.006
	(0.021)	(0.040)	(0.022)	(0.037)	(0.013)
Political position (0 to 1)					
Linear	0.050	0.124	-0.156	-0.040	-0.117
	(0.120)	(0.220)	(0.142)	(0.219)	(0.078)
Squared	-0.001	-0.001	0.003	0.002	0.002
	(0.002)	(0.003)	(0.002)	(0.003)	(0.001)
Educational background					
Knowledge index (0 to 1)	0.130***	-0.139	-0.009	-0.100	-0.026
	(0.037)	(0.071)	(0.045)	(0.067)	(0.024)
Field: STEM	0.066*	0.015	-0.029	0.009	-0.021
	(0.026)	(0.048)	(0.025)	(0.045)	(0.017)
Field: Econ. or business	0.057	-0.058	-0.035	-0.018	-0.018
	(0.037)	(0.061)	(0.031)	(0.061)	(0.020)
Psychological measures					
CRT score (0 to 1)	0.267***	-0.444***	-0.067	-0.304***	-0.082*
	(0.045)	(0.082)	(0.048)	(0.084)	(0.039)
Pseudoscience score (0 to 1)	-0.049	0.244*	-0.058	0.240*	-0.018
	(0.056)	(0.108)	(0.061)	(0.103)	(0.029)
High stakes	0.004	-0.082*	-0.029	-0.011	-0.026*
	(0.018)	(0.036)	(0.021)	(0.033)	(0.012)
Observations	11228	3208	3208	3208	3208
Subjects	401	401	401	401	401

Notes: Column 1 includes three design control dummies (for whether the correct model is associated with the high promise, with the minimal action, and the interaction of the two). Column 1 uses 14 choices for each subject (training and test set); Columns 2 to 5 use 4 choices each. The ORIV-stacked regression doubles these numbers. All regressions exclude 10 subjects who identify as neither male nor female. Knowledge index is the fraction of the following questions a subject can answer correctly: 1. Name of $P(A|B)$, 2. Complete "Correlation does not..." 3. Spell out 'DAG'. Omitted category for gender is male. The omitted category for field of study is 'other.' Political position is the position of the preferred political party according to Jolly et al. (2022), with higher values indicating a more right-wing orientation. Pseudoscience score (Torres et al., 2020) is higher the more an individual believes in pseudoscience. Standard errors in parenthesis, clustered by subject. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

4.5 Individual characteristics

Finally, we examine the effects of subjects' background characteristics. Column 1 of Table 4 reports how these characteristics relate to the probability of choosing the correct model for a given menu in set $\mathcal{M}$. We use OR-IV regression (Gillen et al., 2019) to correct for measurement error in risk and ambiguity preferences, each elicited twice. The specification includes controls for whether the correct model makes the higher or lower promise, whether it recommends the higher or lower investment, and the interaction between these variables. Because stakes have limited effects, we pool across stakes conditions and include an indicator for the High Stakes condition.[39]

Reassuringly, we find that greater background knowledge of statistics and causal inference positively predicts choice of the correct model ($p < 0.01$ in both cases), as does studying a STEM field ($p < 0.05$), and, although insignificantly so, studying economics or business. The effects are sizeable compared to the random choice benchmark of 50%. Other variables have no explanatory power. In particular, and contrary to common expectations, we do not find an effect of political position or political extremity, measured by the square term.This result appears less surprising when considering that both sides of the political spectrum tend to doubt their political opponents' capacity to make rational inferences from observations (*cf. naïve realism*; Griffin and Ross, 1991).

Ex ante, the effects of risk and ambiguity preferences in our setting are ambiguous because of countervailing forces. The risk of selecting an incorrect model can be reduced by investing more effort in checking data charts, yet that effort is itself risky if it is uncertain whether a correct model will be found.[40] Empirically, we find that more risk-averse individuals identify the correct model less often ($p < 0.01$), while ambiguity aversion has no statistically significant effect.

[39] We preregistered complementing this analysis with a multinomial logit-type extension of our main estimates, assuming that the three most common types would receive substantial probability weight. However, the results in Sections 4.1 and 4.3 show a predominant influence of the All Correlations criterion and scattered use of other criteria. The low incidence of types other than those using the All Correlations criterion makes the multinomial logit specification unsuitable.

[40] Risk and ambiguity aversion have no statistically significant effects on time spent per round. The most ambiguity-averse subjects view 1.25 more data charts than the least ambiguity-averse subjects ($p < 0.05$), but risk aversion has no such effect.

Similar regressions for the Utility-Focused Part show that subjects with higher CRT scores ($p < 0.05$) and those in the High Stakes condition ($p < 0.1$) choose the Best-Case Promise alternative significantly less often, while subjects with stronger beliefs in pseudoscience choose it more often ($p < 0.1$). These results align with our interpretation that Best-Case Promise choices in the Spread condition under the State-Constant presentation are unlikely to reflect preferences (column 2). Column 3 supports the interpretation that choices under the Action-Constant presentation do reflect preferences: correlations with cognitive ability measures disappear, while correlations with risk and ambiguity preferences emerge. Results from the Dominance condition (columns 4 and 5) are likewise consistent with this interpretation, indicating that choices in the Action-Constant but not in the State-Constant presentation reflect genuine preferences.

5 Conclusion

In this paper, we have experimentally studied how subjects learn and choose in decision problems for which they lack structural information, such as which variables are related and in what way.

We document two main results. First, subjects display a remarkable ability to discard misspecified models based on qualitative inference obtained by comparing models' correlational implications to the data. Second, when unable or unwilling to choose based on facts, subjects opt for cautious alternatives rather than for the Best-Case Promise alternatives assumed in the Narrative Competition literature. To the extent Best-Case Promise choices occur, they largely reflect a failure of contingent reasoning. That failure is induced by the natural way of presenting the implications of models fitted to data, and hence might also occur outside our laboratory setting. Our results are robust to a threefold increase in the magnitudes of the monetary stakes.

Future research should extend our work in several directions. First, since we focused on settings where subjects can distinguish models only through observational data, we did not include Markov-equivalent models. Future work could examine choices between models that differ only in their interventional, but not observational, implications. Second, empirical settings in which individuals learn

from small samples merit attention, as uncertainty about the data becomes relevant. This is particularly pertinent to the literature on Model Persuasion that is driven by the mechanics of overfitting to small samples.

References

Aina, Chiara and Florian Schneider, "Weighting competing models," 2025.

Ambuehl, Sandro and B Douglas Bernheim, "Social Preferences over Ordinal Outcomes," *American Economic Review*, forthcoming.

Andre, Peter, Ingar Haaland, Christopher Roth, and Johannes Wohlfart, "Narratives about the Macroeconomy," 2023.

Angrisani, Marco, Anya Samek, and Ricardo Serrano-Padial, "Competing Narratives in Action: An Empirical Analysis of Model Adoption Dynamics," *Unpublished*, 2023.

_ , Marco Cipriani, and Antonio Guarino, "Strategic sophistication and trading profits: An experiment with professional traders," *Unpublished*, 2022.

Barron, Kai and Tilman Fries, "Narrative persuasion," *WZB Discussion Paper*, 2023.

Benjamin, Daniel J, "Errors in probabilistic reasoning and judgment biases," *Handbook of Behavioral Economics: Applications and Foundations 1*, 2019, *2*, 69–186.

Bó, Pedro Dal and Guillaume R Fréchette, "The evolution of cooperation in infinitely repeated games: Experimental evidence," *American Economic Review*, 2011, *101* (1), 411–429.

_ and _ , "Strategy choice in the infinitely repeated prisoner's dilemma," *American Economic Review*, 2019, *109* (11), 3929–3952.

Camuffo, Arnaldo, Alfonso Gambardella, and Andrea Pignataro, "Theory-driven strategic management decisions," *CEPR DP 17664v2*, 2023.

Caplin, Andrew, David J. Deming, Søren Leth-Petersen, and Ben Weidmann, "Economic Decision-Making Skill Predicts Income in Two Countries," *NBER Working Paper*, 2023, (31674).

Costa-Gomes, Miguel A and Vincent P Crawford, "Cognition and behavior in two-person guessing games: An experimental study," *American Economic Review*, 2006, *96* (5), 1737–1768.

Coutts, Alexander, Christoph Drobner, Boon Han Koh, and Chris Woolnough, "High Stakes, More Mistakes? Belief Elicitation and Incentives," January 2025. Working paper.

Denzau, Arthur T and Douglass C North, "Shared mental models: ideologies and institutions," *Kyklos*, 1994, *47*, 3–31.

Dimmock, Stephen G, Roy Kouwenberg, Olivia S Mitchell, and Kim Peijnenburg, "Estimating ambiguity preferences and perceptions in multiple prior models: Evidence from the field," *Journal of Risk and Uncertainty*, 2015, *51*, 219–244.

Eckel, Catherine C and Philip J Grossman, "Men, women and risk aversion: Experimental evidence," *Handbook of experimental economics results*, 2008, *1*, 1061–1073.

Eliaz, Kfir and Ran Spiegler, "A model of competing narratives," *American Economic Review*, 2020, *110* (12), 3786–3816.

_ , Simone Galperti, and Ran Spiegler, "False Narratives and Political Mobilization," *arXiv preprint arXiv:2206.12621*, 2022.

Engelmann, Jan, Maël Lebreton, Peter Schwardmann, Joel J van der Weele, and Li-Ang Chang, "Anticipatory anxiety and wishful thinking," *American Economic Review*, forthcoming.

Felin, Teppo and Todd R Zenger, "The theory-based view: Economic actors as theorists," *Strategy Science*, 2017, *2* (4), 258–271.

Finucane, Melissa L and Christina M Gullion, "Developing a tool for measuring the decision-making competence of older adults.," *Psychology and aging*, 2010, *25* (2), 271.

Fox, Craig R and Gülden Ülkümen, "Distinguishing two dimensions of uncertainty," in W. Brun, G. Kirkebøen, and H. Montgomery, eds., *Essays in Judgment and Decision Making*, Oslo: Universitetsforlaget, 2011.

Frechette, Guillaume, Emanuel Vespa, and Sevgi Yuksel, "Extracting Models From Data Sets: An Experiment Using Notes-to-Self," *Unpublished*, 2023.

Frederick, Shane, "Cognitive reflection and decision making," *Journal of Economic Perspectives*, 2005, *19* (4), 25–42.

Gilboa, Itzhak, "Decision under uncertainty: State of the science," *Annual Review of Economics*, 2025, *17*.

Gillen, Ben, Erik Snowberg, and Leeat Yariv, "Experimenting with measurement error: Techniques with applications to the caltech cohort study," *Journal of Political Economy*, 2019, *127* (4), 1826–1863.

Glaser, Markus, Zwetelina Iliewa, and Martin Weber, "Thinking about prices versus thinking about returns in financial markets," *The Journal of Finance*, 2019, *74* (6), 2997–3039.

Griffin, Dale W and Lee Ross, "Subjective construal, social inference, and human misunderstanding," in "Advances in experimental social psychology," Vol. 24, Elsevier, 1991, pp. 319–359.

Griffiths, Thomas L., Nick Chater, and Joshua Tenenbaum, *Bayesian Models of Cognition: Reverse Engineering the Mind*, MIT Press, 2024.

Hanaki, Nobuyuki, Keigo Inukai, Takehito Masuda, and Yuta Shimodaira, "Participants' Characteristics at ISER-Lab in 2020," ISER Discussion Paper 1141, Osaka University, Institute of Social and Economic Research (ISER) September 2021.

Jolly, Seth, Ryan Bakker, Liesbet Hooghe, Gary Marks, Jonathan Polk, Jan Rovny, Marco Steenbergen, and Milada Anna Vachudova, "Chapel Hill expert survey trend file, 1999–2019," *Electoral studies*, 2022, *75*, 102420.

Kendall, Chad W and Constantin Charles, "Causal Narratives," *Unpublished*, 2022.

Klusowski, Joowon, Deborah A Small, and Joseph P Simmons, "Does choice cause an illusion of control?," *Psychological Science*, 2021, *32* (2), 159–172.

Levy, Gilat, Ronny Razin, and Alwyn Young, "Misspecified politics and the recurrence of populism," *American Economic Review*, 2022, *112* (3), 928–962.

Molavi, Pooya, "Macroeconomics with Learning and Misspecification: A General Theory and Applications," 2019.

_ , Alireza Tahbaz-Salehi, and Andrea Vedolin, "Model Complexity, Expectations, and Asset Prices," *NBER working paper*, 2021.

Niederle, Muriel and Emanuel Vespa, "Cognitive Limitations: Failures of Contingent Thinking," *Annual Review of Economics*, 2023, *15*, 307–328.

Olea, José Luis Montiel, Pietro Ortoleva, Mallesh M Pai, and Andrea Prat, "Competing models," *The Quarterly Journal of Economics*, 2022, *137* (4), 2419–2457.

Schumacher, Heiner and Heidi Christina Thysen, "Equilibrium contracts and boundedly rational expectations," *Theoretical Economics*, 2022, *17* (1), 371–414.

Schwartzstein, Joshua and Adi Sunderam, "Using models to persuade," *American Economic Review*, 2021, *111* (1), 276–323.

Shiller, Robert J, "Narrative economics," *American Economic Review*, 2017, *107* (4), 967–1004.

Sloman, Steven, *Causal models: How people think about the world and its alternatives*, Oxford University Press, 2005.

Spiegler, Ran, "Bayesian networks and boundedly rational expectations," *The Quarterly Journal of Economics*, 2016, *131* (3), 1243–1290.

Spirtes, Peter, Clark N Glymour, and Richard Scheines, *Causation, prediction, and search*, MIT press, 2000.

Stefan, Simona and Daniel David, "Recent developments in the experimental investigation of the illusion of control. A meta-analytic review," *Journal of Applied Social Psychology*, 2013, *43* (2), 377–386.

Steyvers, Mark, Joshua B Tenenbaum, Eric-Jan Wagenmakers, and Ben Blum, "Inferring causal networks from observations and interventions," *Cognitive science*, 2003, *27* (3), 453–489.

Thoma, Volker, Elliott White, Asha Panigrahi, Vanessa Strowger, and Irina Anderson, "Good thinking or gut feeling? Cognitive reflection and intuition in traders, bankers and financial non-experts," *PloS one*, 2015, *10* (4), e0123202.

Thomson, Keela S and Daniel M Oppenheimer, "Investigating an alternate form of the cognitive reflection test," *Judgment and Decision making*, 2016, *11* (1), 99–113.

Toplak, Maggie E, Richard F West, and Keith E Stanovich, "Assessing miserly information processing: An expansion of the Cognitive Reflection Test," *Thinking & Reasoning*, 2014, *20* (2), 147–168.

Torres, Marta N, Itxaso Barberia, and Javier Rodríguez-Ferreiro, "Causal illusion as a cognitive basis of pseudoscientific beliefs," *British Journal of Psychology*, 2020, *111* (4), 840–852.

Trautmann, Stefan T and Gijs Van De Kuilen, "Ambiguity attitudes," *The Wiley Blackwell handbook of judgment and decision making*, 2015, *2*, 89–116.

Verma, Thomas S. and Judea Pearl, "Equivalence and Synthesis of Causal Models," in "Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence" Association for Uncertainty in Artificial Intelligence Mountain View, CA 1990, pp. 220–227.

Waldmann, Michael, *The Oxford handbook of causal reasoning*, Oxford University Press, 2017.

Weitzel, Utz, Christoph Huber, Jürgen Huber, Michael Kirchler, Florian Lindner, and Julia Rose, "Bubbles and financial professionals," *The Review of Financial Studies*, 2020, *33* (6), 2659–2696.

Welsh, Matthew Brian and Steve H Begg, "The Cognitive Reflection Test: familiarity and predictive power in professionals," in "Proceedings of the Annual Meeting of the Cognitive Science Society," Vol. 39 2017.

ONLINE APPENDIX

Choice with Competing Models: An Experimental Study

Sandro Ambuehl, Heidi C. Thysen

Table of Contents

A Theory

A.1 Characteristic correlational implications of selected DAGs

Using the notation and language of Pearl (2009), the formal statement of Observation 1 in the case of an arbitrary number of nodes can be stated as follows. This observation follows directly from Theorem 5.2.1 in Pearl (2009).

Observation 2. *Consider a DAG $G = (N, E)$, with $I, J, K \in N$ and $Z \subset N$.*

(i) If $I \to J$, then generically $\text{cov}(I, J) \neq 0$.

(ii) (a) If $Z = \emptyset$ does not d-separate I and J, then generically $\text{cov}(I, J) \neq 0$.

(b) If $Z = \emptyset$ d-separates I and J, then $\text{cov}(I, J) = 0$.

(iii) (a) If $Z = \{K\}$ d-separates I and J, then $\text{cov}(I, J|K) = 0$.

(b) If $Z = \{K\}$ does not d-separate I and J, then generically $\text{cov}(I, J|K) \neq 0$.

A.2 Tools to construct menus that identify criteria distributions

Here, we first construct the action-equivalence classes such that given any DGP, any two DAGs within the same equivalence class recommend the same investment (Subsection A.2.1). Second, we show that for any pair of models we can vary independently whether the model recommending the higher investment also makes the higher or lower promise by adjusting the distribution of investments in the simulated empirical data used for model fitting (Subsection A.2.2). Third, we show that in menus of two linear models, the Worst-Case Promise criterion is equivalent to selecting the model that promises the *lower* payoff if correct (Subsection A.2.3). Throughout we use the notation $A = \sqrt{I}$ (recall that our systems are linear in the square root of I), $X = C_1$, and $Z = C_2$.

Throughout, we will use the following definition.

Definition 4. *Consider a DAG $G = (N, E)$.*

(i) For $I \in N$, $G(I) = \{J \in N \mid (J, I) \in E\}$ is the set of Parents *of node I.*

(ii) For $I, J, K \in N$ we call the triple (I, J, K) a v-collider *if $I, J \in G(K)$, but $I \notin G(J)$, and $J \notin G(I)$.*

Furthermore, we let $V_G(a) = \mathbb{E}_G[Y \mid A = a] - \frac{c}{2}a^2$ denote the expected payoff according to G when action a is implemented.

Throughout, we will make use of the following observation.

Observation 3. *If a linear system of equations with DAG representation G and $G(A) = \emptyset$ is fitted to the data using ordinary least squares, the mean of Y conditional on A is given by*

$$\mathbb{E}_G[Y|A] = \hat{\alpha}^G + \hat{\alpha}^G_A A.$$

When G is consistent with the DGP we let α^*, and α^*_A denote the intercept and slope coefficients. The optimal action recommendation associated with DAG G is the solution to the maximization problem $a^G = \arg\max_a \hat{\alpha}^G + \hat{\alpha}^G_A a - \frac{c}{2}a^2$. Hence, DAG G recommends the action $a^G = \frac{\hat{\alpha}^G}{c}$.

A.2.1 Action equivalence classes

We now define the 15 action-equivalence classes. We characterize the set of DAGs in each equivalence class and report the estimated effect of the action on the bonus for the DAGs in that class. Note that this partition is different from Markov equivalence classes. We make use of the well-known results that for any two variables I and J, the estimated slope coefficient of the regression of J on I is given by $\frac{\text{cov}(J,I)}{\text{var}(I)}$, and that in a regression of a variable J on variables I and K the slope coefficient on I is given by $\frac{\text{cov}(I,J)\text{var}(K) - \text{cov}(I,K)\,\text{cov}(K,J)}{\text{var}(I)\text{var}(K) - \text{cov}(I,K)^2}$. Throughout, variance and covariance operators refer to the DGP.

Class 1 consists of all DAGs that posit no direct or indirect effect of A on Y. Hence, regardless of the DGP, $\hat{\alpha}^G = 0$ for all G in this class.

Class 2 consists of all DAGs with $A \in G(Y)$, and there is no $I \in N$ such that (A, I, Y) is a v-collider. That is, A has a direct influence on Y, and no other variable has a direct influence on Y. While some of the system of linear regressions

represented by a DAG in this class might calculate the total effect of A on Y as the sum of the direct effect of A on Y and the indirect effect of A on Y through one or more of the covariates, the total predicted effect of A on Y is the same. This follows directly from Proposition 2 in Spiegler (2020). For any DAG G in this class,

$$\hat{\alpha}^G = \frac{\text{cov}(A, Y)}{\text{var}(A)}.$$

Class 3 consists of all DAGs with $G(X) = \{A\}$ and $G(Y) = \{X\}$. That is, A does not have an indirect influence on Y, but a direct influence on Y through X, and Z does not (directly or indirectly) influence Y. For any DAG G in this class,

$$\hat{\alpha}^G = \frac{\text{cov}(A, X)}{\text{var}(A)} \frac{\text{cov}(X, Y)}{\text{var}(X)}.$$

Class 4 consists of all DAGs with $G(Z) = \{A\}$ and $G(Y) = \{Z\}$. This class parallels Class 3 with the positions of X and Z switched. For any DAG G in this class,

$$\hat{\alpha}^G = \frac{\text{cov}(A, Z)}{\text{var}(A)} \frac{\text{cov}(Z, Y)}{\text{var}(Z)}.$$

Class 5 consists of the single DAG $G : A \to X \to Z \to Y$. We have

$$\hat{\alpha}^G = \frac{\text{cov}(A, X)}{\text{var}(A)} \frac{\text{cov}(X, Z)}{\text{var}(X)} \frac{\text{cov}(Z, Y)}{\text{var}(Z)}.$$

Class 6 consists of the single DAG $G : A \to Z \to X \to Y$. It parallels Class 5 with the positions of X and Z switched. We have

$$\hat{\alpha}^G = \frac{\text{cov}(A, Z)}{\text{var}(A)} \frac{\text{cov}(Z, X)}{\text{var}(Z)} \frac{\text{cov}(X, Y)}{\text{var}(X)}.$$

Class 7 consists of all DAGs that contain the v-collider (A, X, Y) and no other v-colliders. For any DAG G in this class,

$$\hat{\alpha}^G = \frac{\text{cov}(A, Y)\text{var}(X) - \text{cov}(A, X)\,\text{cov}(X, Y)}{\text{var}(A)\text{var}(X) - \text{cov}(A, X)^2}.$$

Class 8 consists of all DAGs that contain the v-collider (A, Z, Y) and no other v-colliders. It parallels Class 7 with the positions of X and Z switched. For any DAG G in this class,

$$\hat{\alpha}^G = \frac{\text{cov}(A,Y)\text{var}(Z) - \text{cov}(A,Z)\,\text{cov}(Z,Y)}{\text{var}(A)\text{var}(Z) - \text{cov}(A,Z)^2}.$$

Class 9 consists of all DAGs for which $G(Y) = \{A, X, Z\}$ and $A \notin G(X), G(Z)$. For any DAG G in this class,

$$\hat{\alpha}^G = \frac{\text{cov}(AY)(\text{var}(X)\text{var}(Z) - \text{cov}(X,Z)^2)}{\text{var}(A)\text{var}(X)\text{var}(Z) + 2\,\text{cov}(A,X)\,\text{cov}(A,Z)\,\text{cov}(X,Z) - \text{cov}(X,Z)^2\text{var}(A) - \text{cov}(A,X)^2\text{var}(Z) - \text{cov}(A,Z)^2\text{var}(X)}$$
$$- \frac{\text{cov}(A,X)(\text{cov}(X,Y)\text{var}(Z) - \text{cov}(X,Z)\,\text{cov}(Z,Y)) + \text{cov}(A,Z)(\text{cov}(Z,Y)\text{var}(X) - \text{cov}(X,Z)\,\text{cov}(X,Y))}{\text{var}(A)\text{var}(X)\text{var}(Z) + 2\,\text{cov}(A,X)\,\text{cov}(A,Z)\,\text{cov}(X,Z) - \text{cov}(X,Z)^2\text{var}(A) - \text{cov}(A,X)^2\text{var}(Z) - \text{cov}(A,Z)^2\text{var}(X)},$$

which is the slope coefficient on A in the regression of Y on the three regressors A, X, Z.

Class 10 consists of the single DAG $G : A \to X \to Y \leftarrow Z$. We have

$$\hat{\alpha}^G = \frac{\text{cov}(A,X)}{\text{var}(A)} \frac{\text{cov}(X,Y)\text{var}(Z) - \text{cov}(X,Z)\,\text{cov}(Z,Y)}{\text{var}(X)\text{var}(Z) - \text{cov}(X,Z)^2}.$$

Class 11 consists of the single DAG $G : A \to Z \to Y \leftarrow X$. It parallels Class 10 with the positions of X and Z switched. We have

$$\hat{\alpha}^G = \frac{\text{cov}(A,Z)}{\text{var}(A)} \frac{\text{cov}(Z,Y)\text{var}(X) - \text{cov}(X,Z)\,\text{cov}(X,Y)}{\text{var}(X)\text{var}(Z) - \text{cov}(X,Z)^2}.$$

Class 12 consists of the single DAG  . We have

$$\hat{\alpha}^G = \frac{\text{cov}(A,Z)\text{var}(X) - \text{cov}(A,Z)\,\text{cov}(X,Z)}{\text{var}(A)\text{var}(Z) - \text{cov}(A,Z)^2} \frac{\text{cov}(X,Y)}{\text{var}(X)}.$$

Class 13 consists of the single DAG  . We have

$$\hat{\alpha}^G = \frac{\text{cov}(A,X)\text{var}(Z) - \text{cov}(A,X)\,\text{cov}(X,Z)}{\text{var}(A)\text{var}(X) - \text{cov}(A,X)^2} \frac{\text{cov}(Z,Y)}{\text{var}(Z)}.$$

Class 14 consists of the single DAG  . We have

$$\hat{\alpha}^G = \frac{\text{cov}(A,Z)\text{var}(X) - \text{cov}(A,Z)\,\text{cov}(X,Z)}{\text{var}(A)\text{var}(Z) - \text{cov}(A,Z)^2} \frac{\text{cov}(X,Y)\text{var}(Z) - \text{cov}(X,Z)\,\text{cov}(Z,Y)}{\text{var}(X)\text{var}(Z) - \text{cov}(X,Z)^2}.$$

Class 15 consists of the single DAG  . We have

$$\hat{\alpha}^G = \frac{\text{cov}(A,X)\text{var}(Z) - \text{cov}(A,X)\,\text{cov}(X,Z)}{\text{var}(A)\text{var}(X) - \text{cov}(A,X)^2} \frac{\text{cov}(Z,Y)\text{var}(X) - \text{cov}(X,Z)\,\text{cov}(X,Y)}{\text{var}(X)\text{var}(Z) - \text{cov}(X,Z)^2}.$$

This is a comprehensive list of all DAGs we consider for the experiment. We exclude the DAG  because one of its characteristic independence relationships, $A \perp\!\!\!\perp Y|(X,Z)$, involves conditioning on two variables simultaneously, which is not information that we provide to subjects.

A.2.2 Pairwise comparison of promises

We next demonstrate how to select the mean of the action in the DGP to change which of two given models yields the higher promise. To do so, we will make use of lemma 1. An immediate implication of this lemma is that if the action is set to its mean in the data, then any model predicts the same mean outcome. This observation will prove useful to compare the predicted payoffs of the recommended actions across models.

Lemma 1. *Consider a system of linear equations represented by the DAG $G = (N, E)$, where $G(A) = \emptyset$. For every DGP and $I \in N$, we have*

$$\mathbb{E}[\mathbb{E}_G[I \mid A]] = \mathbb{E}[I].$$

Proof. We prove this statement by induction. To anchor the induction, consider any variable $I \in N$ for which $G(I) = \emptyset$. If $I = A$, this holds trivially. If $I \neq A$, then $\mathbb{E}_G[I \mid A] = \mathbb{E}[I]$, since G treats I and A as exogenous variables, and therefore as independent. Hence, $\mathbb{E}[\mathbb{E}_G[I \mid A]] = \mathbb{E}[I]$.

Next, consider any node J and suppose that the induction hypothesis $\mathbb{E}[\mathbb{E}_G[I \mid A]] = \mathbb{E}[I]$ holds for every $I \in G(J)$. Let $\hat{\beta}_{IJ}$ denote the slope coefficient on variable I in the OLS regression of J on all its parents. Then, the constant term

in that regression, $\hat{\beta}_J$ is given by

$$\hat{\beta}_J = \mathbb{E}[J] - \sum_{I \in G(J)} \hat{\beta}_{IJ} \mathbb{E}[I]. \tag{1}$$

Furthermore, applying the conditional expectation operator $E_G[\cdot|A]$ to the regression equation that defines J according to G yields

$$\mathbb{E}_G[J \mid A] = \hat{\beta}_J + \sum_{I \in G(J)} \hat{\beta}_{IJ} \mathbb{E}_G[I \mid A]. \tag{2}$$

Substituting 1 into equation 2, and taking the expectation over the action, we obtain:

$$\mathbb{E}[\mathbb{E}_G[J \mid A]] = \mathbb{E}[J] + \sum_{I \in G(J)} \hat{\beta}_{IJ} (\mathbb{E}[\mathbb{E}_G[I \mid A]] - \mathbb{E}[I])$$

By the induction hypothesis, the term in parentheses is zero. Hence, $\mathbb{E}[\mathbb{E}_G[J \mid A]] = \mathbb{E}[J]$ as was to be shown. □

We can now state the key result, Proposition 1, that the model with the lower action recommendation make the higher promise if and only if the mean action exceeds some threshold.

Proposition 1. *Consider two models, G and G', where $G(A) = G'(A) = \emptyset$. Let a_G and $a_{G'}$ denote the corresponding action recommendations, and suppose $a_G > (<) a_{G'}$. Then $V_G(a^G) \geq V_{G'}(a^{G'})$ if and only if $\mathbb{E}[A] \leq (\geq) \frac{a^G + a^{G'}}{2}$.*

Proof. First, recall that for every DAG, G, we can write the predicted conditional mean of the bonus as a linear function of the action, a, specifically, $\mathbb{E}_G[Y \mid A = a] = \hat{\alpha}^G + \hat{\alpha}^G_A a$. By Lemma 1, we thus have

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}_G[Y \mid A]] = \hat{\alpha}^G + \hat{\alpha}^G_A \mathbb{E}[A]$$
$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid A]] = \alpha^* + \alpha^*_A \mathbb{E}[A].$$

Combining the two equations and solving for $\hat{\alpha}^G$ yields:

$$\hat{\alpha}_G = \alpha^* + (\alpha^*_A - \hat{\alpha}^G_A) \mathbb{E}[A]. \tag{3}$$

We use (3) to express the model's promise as a function of $\mathbb{E}[A]$:

$$\begin{aligned} V_G(a^G) &= \hat{\alpha}^G + \hat{\alpha}_A^G a^G - \frac{c}{2}\left(a^G\right)^2 \\ &= \alpha^* + (\alpha_A^* - \hat{\alpha}_A^G)\mathbb{E}[A] + \hat{\alpha}_A^G a^G - \frac{c}{2}\left(a^G\right)^2 \\ &= \alpha^* + (\alpha_A^* - c \cdot a^G)\mathbb{E}[A] + \frac{c}{2}\left(a^G\right)^2, \end{aligned}$$

where the third equality uses $a^G = \frac{\hat{\alpha}_A^G}{c}$. We use this expression to write the difference between the promises associated with models G and G', respectively, as follows:

$$\begin{aligned} V_G(a^G) - V_{G'}(a^{G'}) = & \quad c \cdot (a^{G'} - a^G)\mathbb{E}[A] + \frac{c}{2}\left(\left(a^G\right)^2 - \left(a^{G'}\right)^2\right) \\ = & \quad c \cdot (a^G - a^{G'})\left(\frac{a^G + a^{G'}}{2} - \mathbb{E}[A]\right). \end{aligned}$$

This concludes the proof. □

A.2.3 The Worst-Case Promise criterion

Here we show that in menus of two linear models, the Worst-Case Promise criterion is equivalent to selecting the model that promises the *lower* payoff if correct.

Proposition 2. *Let $\mathcal{G}$ be the set of available DAGs, with $|\mathcal{G}| = 2$. Then, DAG $G^* \in \mathcal{G}$ implies the lowest promise if and only if*

$$G^* \in \arg\max_{G' \in \mathcal{G}} \min_{G \in \mathcal{G}} V_G(a^{G'}).$$

Proof. Consider two DAGs G and G'. The expected payoff predicted by model G if recommendation $a^{G'}$ is implemented is given by:

$$\begin{aligned} V_G(a^{G'}) &= \hat{\alpha}_G + \hat{\alpha}_A^G a^{G'} - \frac{c}{2}\left(a^{G'}\right)^2 \\ &= \alpha^* + (\alpha_A^* - \hat{\alpha}_A^G)\mathbb{E}[A] + \hat{\alpha}_A^G a^{G'} - \frac{c}{2}\left(a^{G'}\right)^2 \\ &= \alpha^* + (\alpha^* - c \cdot a^G)\mathbb{E}[A] + c \cdot a^G a^{G'} - \frac{c}{2}\left(a^{G'}\right)^2, \end{aligned}$$

where the second equality follows from the steps used in the proof of Lemma 1, and the third equality follows from $a^G = \frac{\hat{\alpha}_A^G}{c}$.

The expected payoff according to G when action recommendation $a^{G'}$ is implemented is higher than the expected payoff according G' when the action recommendation a^G is implemented if and only if

$$V_G(a^{G'}) - V_{G'}(a^G) = c \cdot (a^{G'} - a_A^G)\mathbb{E}[A] + \frac{c}{2}\left(\left(a^G\right)^2 - \left(a^{G'}\right)^2\right) \geq 0,$$

or equivalently,

$$c \cdot (a^G - a^{G'})\left(\frac{a^G + a^{G'}}{2} - \mathbb{E}[A]\right) \geq 0.$$

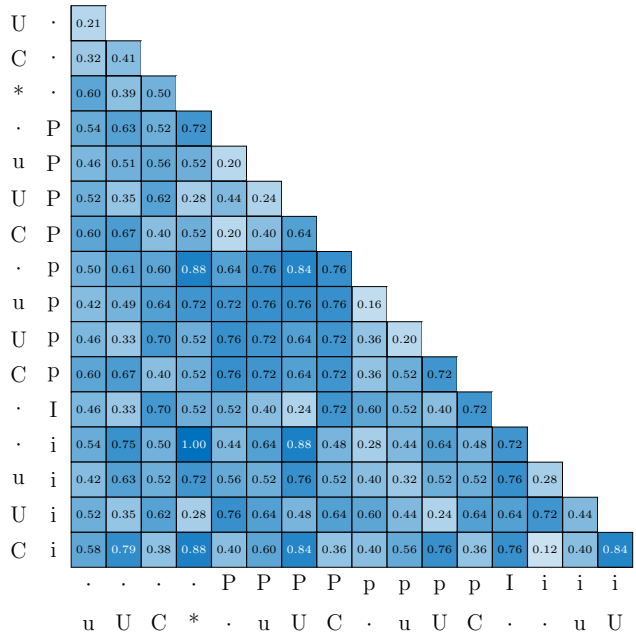
By Lemma 1, $V_G(a^G) > V_{G'}(a^{G'})$ if and only if $c \cdot (a^G - a^{G'})\left(\frac{a^G + a^{G'}}{2} - \mathbb{E}[A]\right) > 0$. Which in turn is equivalent to $V_G(a^{G'}) > V_{G'}(a^G)$. Furthermore, by definition of $a^{G'}$, and $a^{G'} \neq a^G$, we have $V_{G'}(a^{G'}) \geq V_{G'}(a^G)$. Hence, $V_{G'}(a^G) \leq \min\{V_G(a^{G'}), V_{G'}(a^{G'})\}$, this completes the proof. □

B Identification and estimation

B.1 Distance between types

Figure B.1 shows the distance between any pair of types on the choice sets of Experiment 1. We measure the distance between types t and t' as $d(t,t') = (\boldsymbol{M}(0)\boldsymbol{I}_t)\,(\boldsymbol{M}(0)\boldsymbol{I}_{t'})'$ where $\boldsymbol{M}(0)$, as defined in section B.2, is the matrix of theoretically predicted moments when the probability of noisy choices equals $q = 0$ and $\boldsymbol{I}_t$ and $\boldsymbol{I}_{t'}$ are column vectors that have entries one in positions t and t', respectively, and zero everywhere else.

Figure B.1: Distances between types



Notes: The number in each cell and the cell's color indicate the distance between the two types defining that cell. Distances normalized from 0 to 1. Each type is listed as a pair of criteria (Fact-Based, Utility-Based). In each class, a period (·) stands for 'none.' The remaining criteria are encoded as follows. *Data-based*: u: Direct Links, U: Unconditional correlations, C: Conditional correlations. $*$: All correlations. *Utility-based*: P: Best-Case Promise, p: Worst-Case Promise, A: Maximize Investment, a: Minimize Investment.

B.2 GMM estimation

Let n denote the number of types, and let $\boldsymbol{t} = (t_1, \ldots, t_n)$ denote an element of the n-simplex Δ^n that represents the distribution over the types. Each type i is

associated with a matrix T^i with elements $T^i_{c,m}$ that indicates the probability with which type i chooses option c from menu m. Hence, if i makes a unique choice on menu m, then $T^i_{c,m} = 1$ if c is the chosen option and 0 otherwise. If i randomizes, then $T^i_{c,m} = 1/2$.

We use the generalized method of moments to obtain an estimate $\hat{t}$ of the type vector and an estimate $\hat{q}$ of the noise probability. To state the optimization problem formally, and to prove identification of our model, note that the probability that type i chooses option c in menu m when the noise probability is q is given by $\tilde{p}^i_{c,m}(q) = (1-q)T^i_{c,m} + q\frac{1}{2}$. Similarly, $\tilde{p}^i_{(c,m),(c',m')}(q) = (1-q)^2 T^i_{c,m}T^i_{c',m'} + q^2\frac{1}{4} + q(1-q)\frac{1}{2}(T^i_{c,m} + T^i_{c',m'})$ is the probability that type i chooses option c from menu m *and* option c' from menu m'. Given (t, q), our model then predicts first moments $\tilde{p}_{c,m} = \sum_{i=1}^n t_i \tilde{p}^i_{c,m}(q)$, and second moments $\tilde{p}_{(c,m)(c',m')} = \sum_{i=1}^n t_i \tilde{p}^i_{(c,m),(c',m')}(q)$. Note that some of these moments are redundant since (conditional) choice probabilities across all options in a menu must sum to one. We remove redundant moments.

Let $\tilde{p}^i(q)$ denote the column vector of type i's non-redundant first and second moments. Define $\boldsymbol{M}(q) = \big(\tilde{p}^1(q), \ldots, \tilde{p}^n(q)\big)$. Given a type distribution $\boldsymbol{t}$, the vector of theoretically predicted moments is $\boldsymbol{M}(q)\cdot\boldsymbol{t}$. Let $\tilde{\boldsymbol{\mathcal{E}}}$ denote the corresponding empirical moments. Our estimator is then defined as

$$(\hat{\boldsymbol{t}}, \hat{q}) = \arg\min_{\boldsymbol{t},q} \left(\boldsymbol{M}(q)\boldsymbol{t} - \tilde{\boldsymbol{\mathcal{E}}}\right) \boldsymbol{W} \left(\boldsymbol{M}(q)\boldsymbol{t} - \tilde{\boldsymbol{\mathcal{E}}}\right)^{\mathsf{T}} \text{ s.t. } \boldsymbol{t} \in \Delta^n, q \in [0,1] \qquad (4)$$

For the weighting matrix $\boldsymbol{W}$ we use the optimal weighting matrix derived from two-stage feasible GMM.

Regarding identification, note that for a given noise parameter q, the type frequencies are identified only if $\boldsymbol{M}(q)\bar{\boldsymbol{t}} = \boldsymbol{M}(q)\bar{\boldsymbol{t}}'$ implies $\bar{\boldsymbol{t}} = \bar{\boldsymbol{t}}'$, that is, if the nullspace of the linear map $\boldsymbol{M}(q)$ from the type space to the moment space is empty, a condition that we check for our set of menus $\mathcal{M}$, as well as separately for the training and test sets $\mathcal{M}^{train}$ and $\mathcal{M}^{test}$, respectively. To estimate the model with endogenous q, we start the estimation procedure on a grid of initial values for q that spans the unit interval and check the local identification condition at the resulting estimates.

C Experiment design details

Table C.1 shows an overview of the structure of the study, which was coded in Qualtrics and javascript. Here, we list details about each of the stages.

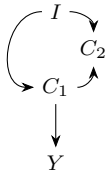
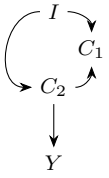
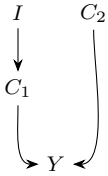
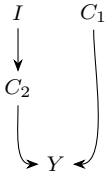
Table C.1: Experiment structure

1. Comprehensive Part
   (a) Instructions and two comprehension checks
   (b) Preliminary rounds
   (c) Main decisions
2. Utility-Focused Part
   (a) Instructions
   (b) Block-randomized decisions, with comprehension check right before
3. Additional decisions
   (a) Ambiguity preference elicitation
   (b) Risk preference elicitation
   (c) Explanation of own decision-making (free response and multiple choice)
   (d) Questions eliciting the understanding of data charts
   (e) Cognitive Response Test
   (f) Pseudoscience scale
   (g) Educational background and demographics

Notes: The experiment proceeds in the order listed. Within each part of section 2, the order of rounds is randomized at the individual level. The two rounds of risk elicitation are also shown in individually randomized order.

Instructions and comprehension check We display all instructions on screen. The entire experiment is in English and was advertised as such. A good command of English is a curricular requirement for all students in our subject pool.

Table C.2: Practice Menus

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Menu	DGP	Competitor	Description	Criterion identifies misspecified model			Model chosen by	
				Direct Links	Uncond. corr.	Cond. corr.	Best-Case Promise	Min. Invest.
P_1			Your Action indirectly influences your Bonus through one of the Counters. The second Counter is a symptom of both your Action and the first Counter.	No	No	Yes	DGP	Comp.
P_2			Your Action indirectly influences your Bonus through one of the counters. The other counter is not influenced by anything, but it has an additional, direct influence on your Bonus.	Yes	Yes	Yes	Comp.	Comp.

Notes: I denotes the investment, referred to as Action in the videos. Column 4 shows the text spoken in the video. In the screens that correspond to Figure 1, the counters are referred to by color. In the case of M_1, for instance, the text for one of the models is "Your Bonus only depends on the red counter. Your Action influences that Counter both directly and through the blue counter.'

We randomize the following display elements: (i) The order in which the list of charts is displayed. For each individual, we randomly select an order of charts displaying unconditional relations that we keep constant for the entire experiment. (ii) Which advisor is on the left of the screen and which is right is randomized in each round for each individual. (iii) The position of models and recommendations within the advisor speech bubbles. This is kept constant for a given subject. For half of the subjects, the models are on top, for the other half the recommendations are on top. (iv) Whether an advisor's promise is listed above or below his claim about how much the subject can expect to earn from following the competing advisor. A subject either sees all or none of the models presented in the opposite order, both in the Comprehensive Part and in the Utility-Focused Part. (v) We randomly redraw the colors of the advisors in each round for each individual to prevent subjects from forming beliefs such as 'the red advisor tends to be correct.'

Practice menus All subjects first completed the two practice menus listed in Table C.2 in random order. Practice menus were not identified as such and could determine payments. As preregistered, we discard choices from practice menus.

Parameters Table C.3 displays the parameters used in each round. Table C.4 displays the resulting model recommendations, promises, and claims about the competing models' recommendation when fit to these DGP, for the High Stakes condition. In the Low Stakes condition, the corresponding amounts are two thirds lower.

Table C.3: DGP parameters

Round	β_A	β_X	β_Y	β_Z	β_{AX}	β_{AY}	β_{AZ}	β_{XY}	β_{XZ}	β_{YZ}	σ_A	σ_X	σ_Y	σ_Z
P_1	0.99	0.00	1.50	0.50	3.00	0.00	-1.90	1.00	0.84	0.00	1.00	2.00	2.00	1.00
P_2	2.01	0.00	0.99	0.50	1.12	0.00	0.00	2.00	0.00	1.00	1.00	1.00	1.00	1.00
M_1	3.46	2.00	5.17	-2.50	-1.70	0.00	2.84	2.10	0.84	0.00	1.00	0.01	6.50	3.70
M_2	1.67	4.00	0.74	-3.50	-1.70	0.00	2.84	2.10	0.84	0.00	1.00	0.01	6.50	3.70
M_3	1.34	0.50	-1.78	0.50	1.30	0.00	0.00	0.00	1.50	1.72	1.00	0.01	4.40	2.76
M_4	1.77	1.00	0.54	-1.20	0.00	3.78	0.00	2.00	2.00	1.40	1.00	1.00	1.00	1.00
M_5	3.52	1.00	-7.23	3.00	0.00	3.75	0.00	2.04	2.00	1.40	1.00	1.00	1.00	1.00
M_6	1.07	0.05	1.26	0.00	0.00	3.08	-2.00	2.00	0.00	2.39	1.00	0.90	0.03	0.19
M_7	2.86	2.00	-7.93	3.00	0.00	3.09	-2.00	2.00	0.00	2.39	1.00	0.90	0.03	0.29
M_8	1.59	-1.00	2.93	0.50	1.60	0.00	0.00	-3.15	2.00	2.00	1.00	0.50	1.00	3.00
M_9	3.38	-1.00	16.93	-5.50	1.60	0.00	0.00	-3.15	2.00	2.00	1.00	0.50	1.00	3.00
M_{10}	1.60	0.00	-0.52	1.00	3.80	0.00	0.00	0.95	0.00	0.87	1.00	0.90	1.70	4.54
M_{11}	1.08	-0.59	2.86	0.00	0.00	3.09	0.00	2.77	0.00	1.00	1.00	1.00	0.01	1.72
M_{12}	2.77	1.00	-3.93	1.00	0.00	3.00	0.00	3.43	0.00	1.00	1.00	1.00	0.10	2.00
M_{13}	0.25	1.00	0.37	0.50	0.75	0.00	0.00	3.00	0.20	0.39	1.00	0.31	0.75	1.25
M_{14}	1.65	0.50	-0.41	0.50	1.05	0.00	0.00	2.00	0.20	2.00	1.00	1.00	1.00	1.00
W_1	0.23	1.00	2.43	1.00	0.00	2.49	3.00	1.00	0.00	0.00	1.00	1.00	1.00	1.00
W_2	4.02	-2.72	-0.07	0.00	1.31	0.00	0.00	1.53	2.12	1.53	1.00	0.01	0.43	0.07
W_3	1.50	-1.45	0.85	1.00	3.80	0.00	0.00	1.00	0.00	0.40	1.00	1.00	5.20	0.74
W_4	2.78	-1.00	1.28	1.00	1.67	0.00	0.00	1.80	0.45	0.00	1.00	1.00	1.00	1.50

Notes: Each model in our setting is a system of linear Gaussian equations. For any variable i, β_i denotes the constant term in the equation corresponding to endogenous variable i, and σ_i is the standard deviation of the corresponding error term. For any endogenous variable i that depends on some other variable j, β_{ij} is the slope coefficient on variable j in the equation corresponding to endogenous variable i.

Table C.4: Models' recommendations, promises, and claims about the competitor's recommendation

Round	Model 1			Model 2		
	Recommendation	Promise	Claim about competitor	Recommendation	Promise	Claim about competitor
P_1	54.00	72.00	41.89	3.47	48.02	17.92
P_2	30.11	48.00	17.89	0.00	72.00	41.90
M_1	12.48	72.00	41.87	81.40	48.01	17.88
M_2	12.48	48.00	17.87	81.40	72.01	41.88
M_3	67.57	72.01	42.01	7.52	48.02	18.02
M_4	85.73	72.03	41.95	14.25	48.05	17.97
M_5	84.38	48.00	17.99	13.74	71.98	41.96
M_6	56.92	72.04	41.98	4.25	48.03	17.98
M_7	57.21	48.02	17.90	4.31	72.00	41.89
M_8	11.10	48.01	17.87	77.81	72.01	41.88
M_9	11.10	72.06	41.92	77.81	48.00	17.87
M_{10}	78.19	72.00	41.89	11.26	48.01	17.90
M_{11}	57.44	72.00	41.99	4.41	48.04	18.03
M_{12}	54.00	47.98	17.89	3.47	72.02	41.93
M_{13}	30.38	76.67	46.29	0.00	56.50	8.94
M_{14}	26.46	48.00	21.54	0.00	72.03	34.42
W_1	0.00	48.06	17.92	30.13	72.00	41.87
W_2	24.10	72.03	41.85	108.23	48.05	17.87
W_3	74.05	72.03	42.02	9.78	48.01	18.01
W_4	54.22	48.04	17.89	3.51	72.05	41.90

Notes: This table shows the numbers for the High Stakes condition. In the Low Stakes condition, all numbers are two thirds lower.

Risk preference elicitation There are two rounds. In each subjects chose one of six lotteries each of which offers a 2/3 chance of obtaining a higher amount and a 1/3 chance of obtaining a lower amount. In the first round, the amounts in Swiss Francs in the Low Stakes treatment are (15, 15), (16, 14), (20, 10), (24, 6), (28, 2). In the second round, they are (18.75, 18.75), (20, 17.5), (25, 12.5), (30, 7.5), (35, 2.5). In the High Stakes treatment, all amounts are tripled. We randomize whether the lotteries are ordered from safe to risky or from risky to safe.

Demographics and other characteristics In addition to the characteristics listed in Section 3.3, we also elicit the following characteristics: (i) native language, (ii) country of origin, (iii) age, (iv) degree level the subject is working towards, (v) monthly spending, (vi) religiosity (vii) eligibility to vote in political elections in Switzerland, (viii) how much the person agrees with the political party to which they are closest

In terms of educational background, we elicit the institution and faculty at which the subject is enrolled in their main field of study. For the purpose of analysis, we will classify these institutions into STEM, business/economics, and other.

D Supplementary Analysis

D.1 Summary statistics

Table D.5 shows summary statistics of our samples. Our sample skews slightly female and to the political left, as is common for student subject pools. Slightly under half of our subjects are from German-speaking Europe, and a bit over a quarter are from Asia.

D.2 Order effects

Panel A of Figure D.2 plots the median response time against the position at which the subject made the corresponding decision. Subjects take substantially longer on the first decision, presumably to familiarize themselves with the interface. While response times decline across the entire experiment, this decline appears to reflect learning rather than decreased attention, as Panel B shows. For each decision position, it plots the fraction of subjects who viewed at least one data chart. Approximately 85% of subjects do so in any given round.

Figure D.2: Order effects

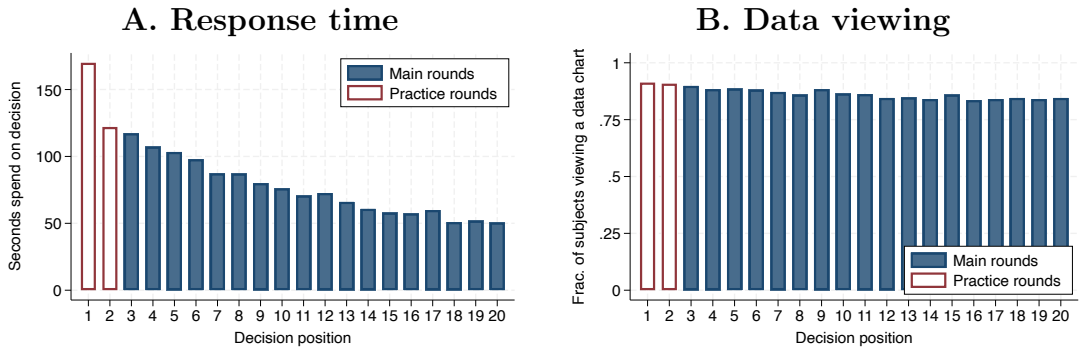


Table D.5: Subject characteristics

Male	0.461
Female	0.539
Age	24.5
Place of origin	
German-speaking Europe	0.225
Non-German-speaking Europe	0.068
Asia	0.135
Other	0.572
Current degree program	
BA	0.192
MA	0.263
PhD and MD/JD	0.029
Not working toward a degree	0.025
Field of study	
STEM	0.643
Economics or business	0.118
Other field	0.219
Statistical knowledge	
Can name $P(A\|B)$	0.246
Can complete "Correlation does not..."	0.536
Can spell out DAG	0.077
Took class on causal inference	0.225
Psychological measures	
CRT score (0–7)	5.072
Pseudoscience score (20–100)	59.560
Religiosity (1–5)	1.696
Closest political party	
SVP	0.028
FDP	0.066
BDP	0.031
CVP	0.045
GLP	0.117
SP	0.130
Green	0.058
PdA	0.033
Subjects	414

Notes: Political parties are listed in order of overall stance on the political spectrum, beginning with most conservative. CVP is the center party.

D.3 Best-fitting types

Table D.6 lists the estimated frequencies of all types estimated on the union of training and test set, along with heteroscedasticity-robust standard errors, separately for the Low and High Stakes treatments.

Use a width-based resize on the two panels. Notes stay unscaled.

Table D.6: Most common types

A. Low stakes

Criterion											Freq.	s.e.
Utility-based					Fact-based							
None	High Promise	Low Promise	High Action	Low Action	None	Direct Links	Unconditonal	Conditional	All			
○									●		0.654	(0.049)
○						●					0.056	(0.026)
		●					●				0.047	(0.012)
○								●			0.046	(0.066)
				●	○						0.038	(0.007)
		●			○						0.037	(0.009)
		●						●			0.029	(0.011)
○							●				0.027	(0.035)
	●				○						0.025	(0.010)
	●					●					0.020	(0.009)
				●		●					0.012	(0.007)
			●		○						0.009	(0.010)
				●			●				0.000	(0.017)
				●				●			0.000	(0.019)
	●						●				0.000	(0.015)
		●				●					0.000	(0.008)
	●							●			0.000	(0.007)

B. High stakes

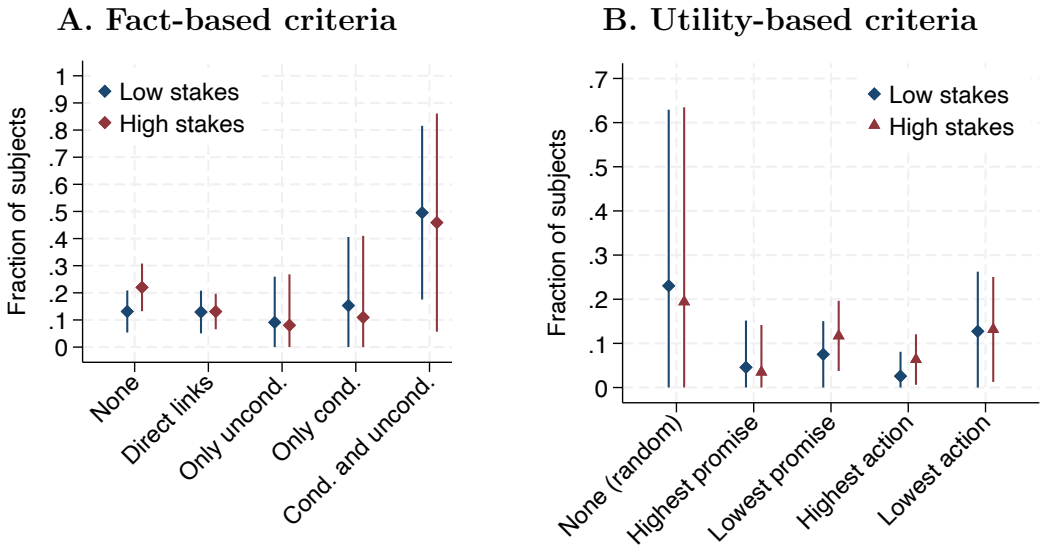
Criterion										Freq.	s.e.
Utility					Factual						
None	High Promise	Low Promise	High Action	Low Action	None	Direct Links	Unconditonal	Conditional	All		
○									●	0.569	(0.048)
				●	○					0.069	(0.010)
		●					●			0.057	(0.012)
○						●				0.047	(0.019)
			●		○					0.047	(0.010)
		●			○					0.043	(0.011)
○								●		0.036	(0.062)
		●						●		0.034	(0.011)
○							●			0.032	(0.036)
	●				○					0.027	(0.008)
	●					●				0.019	(0.008)
				●				●		0.012	(0.019)
		●				●				0.004	(0.007)
				●		●				0.003	(0.005)
				●			●			0.000	(0.012)
	●							●		0.000	(0.007)
	●						●			0.000	(0.011)

Notes: The symbol ● indicates that the corresponding criterion is being used, ○ indicates that no criterion from the corresponding class is being applied. The top row contains only a single symbol because the Conditional Correlations criterion prevents the identification of structure- and advice-based criteria. Heteroskedasticity-robust standard errors in parentheses.

D.4 Distribution of decision criteria

Figure D.3 shows the distribution of decision criteria estimated only on the training set, separately for the low and high stakes conditions.

Figure D.3: Distribution of decision criteria on the training set, split by stakes



Notes: Whiskers show 95%-confidence intervals, truncated at 0. *Panel A: Panel B:* Estimates of advice-based criteria in Experiment 1 are shown conditional on not using the Conditional Correlations criterion.

D.5 Utility-Focused Part

Table D.7 shows choice frequencies and statistical tests for differences across treatment conditions, pooling over the presentation-order condition.

We complement this aggregate analyis with individual-level analysis to ensure the aggregates do not mask important individual-level heterogeneity. To do so, we define the following 12 types. A subject seeks to make either maxmax or maxmin choices in the Model frame and may have the same or the opposite objective in the Gamble frame. She might successfully choose according to her preferences throughout, she might be confused in the state-constant (but not in the action-constant) presentation mode and select the utility-minimizing option there, or she might be confused in the action-constant (but not in the state-constant) presentation mode. For rounds in which one alternative dominates another, both maxmax and maxmin decision makers prefer that alternative. We assign each subject to the type whose choices coincide with that subjects' choices on the largest number of rounds. In case a subjects' choices deviate from multiple different types by the same number of profiles, we calculate type frequencies by assigning equal probability mass to each these multiple types.

Table D.7: Utility-Focused Part: Econometric Tests

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Spread condition		**Dominance condition**		Difference between columns			
Dependent variable	Maxmax choice		Dominant choice		*p*-values			
Stakes	Low	High	Low	High	1-2	3-4	1-3	2-4
Model frame								
Levels								
State constant	0.357***	0.321***	0.762***	0.716***	0.393	0.236	0.000	0.000
	(0.030)	(0.030)	(0.026)	(0.029)				
Action constant	0.095***	0.064***	0.962***	0.978***	0.164	0.264	0.000	0.000
	(0.016)	(0.016)	(0.011)	(0.009)				
Difference	-0.262***	-0.257***	0.200***	0.262***	0.914	0.124	0.000	0.000
	(0.029)	(0.031)	(0.028)	(0.029)				
Gamble frame								
Levels								
State constant	0.379***	0.287***	0.724***	0.745***	0.027	0.589	0.000	0.000
	(0.030)	(0.028)	(0.028)	(0.028)				
Action constant	0.133***	0.081***	0.952***	0.985***	0.047	0.023	0.000	0.000
	(0.020)	(0.017)	(0.013)	(0.006)				
Difference	-0.224***	-0.240***	0.190***	0.270***	0.711	0.049	0.000	0.000
	(0.032)	(0.031)	(0.028)	(0.029)				
p-values: Model vs. Gamble								
State Constant	0.415	0.191	0.161	0.294				
Action Constant	0.074	0.309	0.481	0.469				
Subjects	210	204	210	204				
Observations	1680	1632	1680	1632				

The modal subject, 62.7%, makes maxmin choices throughout and are not systematically confused by any frame (65.2% and 59.3% of subjects in the low and high-stakes conditions, respectively). The next biggest group of subjects, 21.2% make maxmin choices throughout, but are systematically confused by the state-constant frame where they inadvertently make maxmax choices (20.2% and 22.2$ in the low and high stakes conditions, respectively). Of the remaining subjects, 11.5% are approximately uniformly distributed across the remaining types that do not make systematic mistakes, and merely 5.1% are assigned to types that make systematic mistakes but have objectives other than minmin in both frames.

E Experiment instructions

Due to page restrictions, the complete instructions are available online at https://narrativesstudy.s3.us-east-2.amazonaws.com/instructions.pdf

References

Pearl, Judea, *Causality*, Cambridge University Press, 2009.

Spiegler, Ran, "Behavioral implications of causal misperceptions," *Annual Review of Economics*, 2020, *12*, 81–106.