Learning with Mental Models: the case of overestimation

Jimena Galindo

October 15, 2023

Abstract

I design a framework and a laboratory experiment that allows for the comparison of multiple theories of misspecified learning. I focus on a framework with endogenous information and a data-generating process ruled by two fundamentals: an ego-relevant parameter and a state. Within this framework I study three forces that can lead to misspecified beliefs: initial misspecifications, learning traps and biased updating. I find that biased updating is the main driver of misspecified beliefs in the lab. In addition, I vary the degree of ego-relevance of the parameter by introducing a stereotype treatment. The data is consistent with biased updating in both cases but for potentially different reasons: when learning about themselves, subjects attribute successes to their own ability and failures to luck. Instead, in the stereotype treatment, they compensate for initial negative biases by over-attributing positive signals to the ability of others. This translates into similar observed choices but different dynamics in beliefs.

1 Introduction

A growing body of literature in economics explores how people develop incorrect beliefs about fundamentals. Most of this research centers on scenarios where agents passively observe the world, and incorrectly integrate the received information into their beliefs.[1] However, many real-world situations cast agents as active participants in the generation of information. In these cases, the information they observe is influenced by their actions and subsequent behavior is in turn determined by how they incorporate the information into their beliefs.

As an example consider a student who needs to decide how much effort to put into studying for an exam. Their decision will depend on two factors: their belief about their intrinsic ability, and their belief about how difficult the exam will be. The outcome they observe will be affected by how much they decide to study. Imagine that the student puts in a moderate amount of effort and gets a surprisingly good grade. Did they get a good grade because they are smarter than they thought? Or because the exam was easier than they had anticipated? Their future exam preparation strategy will depend fundamentally on which line of reasoning they take. This feedback loop is referred to as an *endogenous information process* and is at the center of the forces I study.

To understand what the main forces at play are, I compare a set of theories that model learning in settings with endogenous information, and which can rationalize the persistence of misspecified beliefs. I develop a unifying framework that nests multiple theories of learning and generates testable predictions for each of them. Then I cast this framework in a laboratory experiment and test the predictions to identify which of the theories are consistent with the behavior I observe in the lab. The experiment features an agent who needs to learn two parameters: one that pertains their own characteristics (an *ego-relevant* parameter), and an exogenously determined state—in the context of the student preparing for exams, the two parameters would be their intrinsic ability and the difficulty of the exam.

Misspecified beliefs about an ego-relevant parameter are often referred to as over-

[1] See Benjamin [2019] for a review of the literature on errors of probabilistic thinking.

confidence (or underconfidence) and have been documented by behavioral scientists[2] and economists[3] in a variety of settings. Oster et al. [2013] find that subjects who are at high risk of having Huntington's disease overestimate their probability of being healthy and make retirement decisions as if they were healthy. Hoffman and Burks [2020] show that workers overestimate the quality of their match to their current employment and are unlikely to look for other opportunities. Camerer and Lovallo [1999] find that entrepreneurs are overconfident about the quality of their enterprise, which leads to excessive entry and early exit from markets.

In all of these examples, holding an incorrect belief about a fundamental leads to suboptimal choices with potentially high costs. In spite of the abundance of evidence, the scope of the existing research is limited in terms of the frameworks it considers. Most of the experimental evidence documenting the bias is collected in settings where subject are passive learners.[4] They observe a noisy signal and report their beliefs over subsequent rounds.[5] Although these studies provide important insights, they are not flexible enough to incorporate the richer theories that have been proposed more recently. In particular, they do not allow the study of endogenous learning or for learning about multiple parameters at once.[6]

In my experiment, I move away from the standard framework of passive learning to analyze a richer set of learning mechanisms. In particular, the interaction between the two parameters together with an endogenous information process gives rise to three possible mechanisms that allow for the persistence of incorrect beliefs: the presence of learning traps, incorrect initial beliefs, and misattribution bias. The theories that I consider incorporate different combinations of these mechanisms.

When the setting features learning traps, even an agent who incorporates all information

[2]Kelley and Michela [1980] provides a review of the psychology literature.

[3]See Benjamin [2019] for a review of the literature in economics.

[4]Götte and Kozakiewicz [2022] and Ozyilmaz [2022] are exceptions that study settings with endogenous information processes.

[5]Bracha and Brown [2012] and Möbius et al. [2022] are some examples.

[6]Coutts et al. [2020] studies an environment with an ego-relevant parameter and an exogenous state but does not incorporate the endogenous information process.

correctly, may fall into learning traps as outlined by Hestermann and Yaouanq [2021]. These traps are characterized by a combination of an incorrect belief and an optimal action which produce information that confirms the incorrect belief. Once an agent falls into a trap, the belief will be stable and even with a correctly specified model of learning, they will not be able to abandon their misspecified beliefs. If, the agent is dogmatic about their initial belief, Heidhues et al. [2018] show that they will inevitably fall into a trap and thus will be able to rationalize and sustain their initial misspecification belief.[7]

Ba [2023] moves away from dogmatism and endows the agents with a mechanism through which they can abandon incorrect beliefs; this allows them to avoid falling into learning traps. To do so agents perform Bayesian hypothesis tests that evaluate which is the more likely parameter out of two possibilities. By doing this, she characterizes the set of situations in which, even agents who consider alternative paradigms, may become trapped.[8]

Lastly, misattribution bias is the more classical explanation and has been widely studied in behavioral science.[9] Agents who suffer from misattribution bias will attribute successes to their own ability—the ego-relevant parameter—and failures to bad luck—the state. Under this model of learning, even an agent who initially has a correct initial belief may become overconfident if they observe a sequence of successes. In this case, the main driver of the bias is not an initial misspecification or the presence of learning traps, it is the updating procedure itself.[10]

These theories provide the main building blocks for a simplified framework that can be directly implemented in a laboratory experiment. In the experiment subjects make choices and receive feedback that depends on their own ability, an exogenous parameter and the choice they made. I track their choices as well as their beliefs about their own ability. The goal is to identify which of the 3 forces—the presence of learning traps, misspecified initial

[7] Götte and Kozakiewicz [2022] study the case of agents with dogmatic initial beliefs in a laboratory experiment.

[8] A similar mechanism is proposed by Schwartzstein and Sunderam [2021] in a setting with persuasion.

[9] See Kelley and Michela [1980] for a review.

[10] A more general framework that can be used to model this bias has also been proposed by Brunnermeier et al. [2005] and empirically studied by Bracha and Brown [2012].

beliefs, or misatribution bias—better explains the observed behavior. To determine the fit of the models, I compare the behavior predicted by each theory to the benchmark given by the fully Bayesian updating procedure.

I also study whether the learning mechanism is inherently linked to the ego-relevance of the parameters or if it is a more general phenomenon. I vary the degree of ego-relevance by introducing a treatment in which subjects learn about the ability of another participant. In this treatment, the participants know only the gender and nationality of the other, and thus can induce stereotypes—a different type of misspecification.

If correct learning about the parameters happens at higher rates in the stereotype treatment, it would suggest that the bias is intrinsically linked to the ego-relevance of the parameter. In contrast, if similar biased behavior arises in both treatments, it is more likely that the main driver of these types of misspecified beliefs is the updating procedure itself or the endogenous information process.

Although some agents do fall into learning traps, I find that the behavior of most subjects is better explained by misattribution bias: good news are treated as signaling high ability, while bad news are attributed to a low state. I also find that misattribution is no more prevalent in the ego-relevant than in the stereotype condition. This suggests that the main driver of the misspecification is the updating procedure; however, the underlying mechanism by which the bias is generated may be different in both treatments: While in the ego-relevant condition subjects prefer to hold themselves in high esteem, in the stereotype condition updating seems to be driven by some sort of bias overcorrection—when subjects realize that they underestimated the ability of another participant based on their gender and nationality, they compensate by overestimating their ability.

Finally, I estimate the structural parameters of the models to study model-heterogeneity in the sample. I find that even at an individual level the behavior is better explained by a general model of misattribution bias for most subjects. There is a smaller group of subjects that can be better explained through dogmatic beliefs and hypothesis testing and none of

them behave in line with the fully Bayesian benchmark.

In what follows I first discuss the theoretical framework and the predictions of each of the theories. Then I introduce a unifying example and my hypotheses. In section 4 I describe the experimental design and in section 5 I present the data and the results. Section 6 outlines the estimation of the parameters and the model fit analysis.

2 Theoretical Framework

The theories that I consider make predictions in two distinct frameworks. I first detail each theory within their original framework and then develop a unifying example that allows me to compare the predictions of all of them. In framework 1 I consider a setting in which the an agent observes a continuous output and infers the underlying state. In framework 2, the agent observes a binary outcome and infers the underlying probability of success. In both frameworks, I consider two nested theories of learning.

2.1 Framework 1

An agent is of type $\theta \in \Theta$ and faces an unknown exogenous state ω drawn from some density f over Ω. The agent knows the distribution of ω but not its realized value. His belief about the state, $p_0(\omega)$, coincide with the true distribution, f; and his belief about the type is $p_0(\theta)$. Let the agent's true type be θ^*, and the realized state be ω^*.

An agent's belief about the type is *misspecified* if it assigns probability zero to their true type. Furthermore, the agent is *dogmatic* if he holds a degenerate belief that places probability one on being of type $\hat{\theta}$. An agent can be dogmatic and misspecified, in which case $\hat{\theta} \neq \theta^*$ and $p_0(\hat{\theta}) = 1$.

The agent chooses an action $a \in A$ and observes a noisy outcome h. The outcome is a function of the agent's type, the state, and the action. In particular $h = h(\theta^*, \omega^*, a) + \varepsilon$ with $h(\cdot)$ increasing in both θ^* and ω^*, and such that conditional on a pair of parameters (θ, ω),

there is a unique optimal action. additionally, $\varepsilon \sim N(0, \sigma)$ is noise in the output.

After observing the outcome, the agent updates his beliefs about θ and ω using some algorithm and moves on to the next period. He repeats this process infinitely many times. I make the simplifying assumption that the agent is myopic and chooses the action that maximizes the payoff in each period. This assumption simplifies the analysis and plays a role in whether an agent who updates their beliefs using Bayes rule would learn the truth or not, however, for the main theory discussed in this section, the results hold even when relaxing this assumption. The behavior of a forward-looking agent is further discussed in the conclusion.

A key notion in this setting is that of a self-defeating equilibrium[11]. A *self-defeating equilibrium* is a belief and action pair such that the agent's belief about their type is misspecified and the outcome generated by the action is consistent with the misspecified belief. This means that the average outcome under the true type and the true state equals the average output the agent expects under the misspecified belief. The agent's belief is said to be *stable* when this happens.

Within this framework, I consider two nested theories of belief updating: the first one is a dogmatic modeler from Heidhues et al. [2018], The second one is a switcher, as in Ba [2023]. The dogmatic modeler can be seen as a switcher with an infinitely sticky initial belief so the two theories are nested. However, outside of the limiting case, they produce different predictions about the agent's behavior. I discuss each of them separately in what follows.

2.1.1 The Dogmatic Modeler

Take the action space to be the same as the state space Ω and $\Omega = A = [\underline{a}, \infty)$.

A dogmatic agent does not update their beliefs about θ, instead, he holds a degenerate belief that places probability on $\hat{\theta}$, which is potentially misspecified. In this case, no matter how much information he gathers against being of type $\hat{\theta}$, he will not update his beliefs about

[11]This notion is an adaptation of the Berk-Nash equilibrium in Esponda and Pouzo [2016] to this setting with only one agent

it. Any discrepancies between the observed outcomes and his believed type are incorporated using the Bayes rule to update his beliefs about ω. Heidhues et al. [2018] show that, under certain assumptions on the per-period utility,[12] a dogmatic modeler will inevitably fall into a self-defeating equilibrium. The equilibrium will be such that the outcomes they observe reinforce their belief on ω in such a way that as $t \to \infty$ the agent will be sure that the state is some $\omega^{'}$ consistent with their believed type and the observed data. In other words, they will be in a self-defeating equilibrium with a stable belief that places probability one on the incorrect parameters $(\hat{\theta}, \omega_\infty)$

The mechanism by which the dogmatic agent falls into the self-defeating equilibrium is the following: Suppose the agent holds the misspecified belief that they are type $\hat{\theta} > \theta^*$. For any prior over ω, the agent will be disappointed by the outcome. He expected a gain of $h(\hat{\theta}, \mathbb{E}(\omega), a)$ but instead observes $h(\hat{\theta}, \omega^*, a)$. There are two possible sources for the disappointment: the first is that the realized state is lower than the expected state; the second source is that the agent is of type θ^* and therefore, for all possible states, his gain will be lower than what he expected. Because the agent is dogmatic, he will not update his beliefs about θ and as a consequence will attribute the disappointment to the state being lower than expected. He will continue to update in this way until he converges to a belief about ω that is stable. Such a belief will explain the observed utility perfectly and allow the agent to rationalize his dogmatic belief about θ. Under the assumptions of Heidhues et al. [2018], there is a unique value of ω at which the belief is stable, I will refer to such value as ω_∞. This mechanism is illustrated in Example 1.

Example 1: *Set $A = \Omega = [\underline{a}, \infty)$ and consider a student with intrinsic ability $\theta^* \geq 0$ who faces a grading procedure ω^* that is unknown to them. The student knows that a higher ω^* is more likely to yield a higher grade. In particular, she knows the grade is given by $(\theta^* + a)\omega^*$.*

[12] The assumptions are that u is twice continuously differentiable with: (i)$u_{aa} < 0$ and $u_a(\underline{a}, \theta, \omega) > 0 > u_a(\bar{a}, \theta, \omega)$, (ii) $u_\theta, u_\omega > 0$ and (iii) $u_{a\theta} < 0$ and $u_{a\omega} > 0$. Where $\underline{a}$ is the minimal action and $\bar{a}$ is the maximal one. The direction of the derivatives is a normalization and the results would hold even when the signs are reversed.

The student must choose an effort level a, which determines her grade. For whatever the chosen effort level, the student must pay a cost $c(a) = \frac{1}{2}a^2$ and observes a grade of $(\theta^* + a)\omega^*$. She repeats this process for infinitely many periods. Assume also that the student's prior is such that $\mathbb{E}[\omega] = \omega^*$ and she is dogmatic about being of type $\hat{\theta} > \theta^*$.[13] *Therefore, the student's payoff in period t is given by*

$$u_t(a_t; \theta^*, \omega^*) = (\theta^* + a_t)\omega^* - \frac{1}{2}a^2 + \varepsilon_t \tag{1}$$

Under this specification, the myopic optimal effort level is $a_t^* = \omega^*$. Because the student does not know ω^*, he will choose $a_t = \mathbb{E}_t(\omega)$ where the expectation is taken with respect to the student's belief at the beginning of period t. If she does not revise his effort choice for k periods, she will receive an average utility of $(\theta^* + a_t^*)\omega^* - \frac{1}{2}a_t^{*2}$ but he was expecting an average utility of $(\hat{\theta} + a_t^*)\omega^* - \frac{1}{2}a_t^{*2}$. In response, she will apply Bayes rule to update his beliefs about ω to get the posterior belief with $\mathbb{E}_{t+k}[\omega] = \frac{(\theta^* + \omega^*)\omega^*}{\hat{\theta} + \omega^*}$ which is lower than the initial belief. This will cause the student to choose a lower effort at $t + k$. As a result, she will again receive an average utility that is lower than what he expected which will cause her belief to drift further down. This process will continue until the average utility equals her expected utility under the dogmatic belief that assigns probability 1 to $\hat{\theta}$. At that point, the student will have reached a self-defeating equilibrium and she will continue to choose sub-optimal effort forever.

Although the model of a dogmatic modeler rationalizes the prevalence of overconfident (underconfident) beliefs, the assumption that the agent has a degenerate belief and no mechanism through which he can update such belief is very restrictive. An alternative approach is proposed by Ba [2023]. She proposes an extension of the dogmatic agent who is able to swithch from one dogmatic belief to another. By doing so, the agent can avoid the self-defeating equilibria and end up being dogmatic and correctly specified.

[13] The example is illustrated for an overconfident student but the results are symmetric for a dogmatic student who initially places probability one on some $\tilde{\theta} < \theta^*$.

2.1.2 The Switcher

An agent is a *switcher* if they behave as a dogmatic, but are willing to entertain the possibility that they are of a different type. In particular, when they start off as a misspecified dogmatic, they are willing to switch to a different dogmatic belief if the data is convincing enough. Their prior is still degenerate and assigns probability one to a particular type, and zero to all other types. This means that a Bayesian update on θ does not change their beliefs about the type. However, they are willing to entertain two such beliefs and have a mechanism by which they decide which belief to adopt at any period t.

In order to abandon their initial dogmatic belief, the agent needs to observe a sequence of outcomes that are sufficiently unlikely to have happened if they were of the type they initially believed. In order to evaluate if the evidence is convincing enough, they keep track of the likelihood that each of the possible types generated the data. If the likelihood ratio is sufficiently large, the agent will switch to the alternative and behave as if they are dogmatic about the new type.

In particular, for an agent that starts with a dogmatic belief that they are of type $\hat{\theta}$ but is willing to consider the alternative explanation that they are of type $\tilde{\theta}$, the agent will switch to the alternative if:

$$\frac{p[h^t|\tilde{\theta}]}{p[h^t|\hat{\theta}]} > \alpha \geq 1$$

Where h^t is the history of outcomes up to time t and α is the switching threshold.[14] By keeping track of the likelihood ratio, the agent can perform a *Bayesian hypothesis test* and adopt the Dogmatic belief that best fits the data.[15]

By allowing the agent to keep track of the likelihoods and switching to an alternative type, the switcher can avoid the self-confirming equilibria. However, if the prior belief on

[14]Notice that if $\alpha \to \infty$, the behavior of the switcher will be indistinguishable from that of the Dogmatic modeler. In this sense, the switcher is a generalization of the dogmatic type.

[15]In a related problem Schwartzstein and Sunderam [2021] proposes a similar updating procedure which relies on the Bayesian hypothesis test. However, in their model there is a sender who optimally chooses to propose a model that fits the data

omega is sufficiently tight around a self-defeating equilibrium, the switcher might look identical to the dogmatic even in a case where α is not too large. This happens because under the agent's prior, the likelihood ratio is unlikely to grow as fast as it is needed to escape the self-defeating equilibrium. In such situations, we say that the misspecified belief is persistent.

2.2 Framework 2

As in framework 1, the agent is of some type $\theta^* \in \Theta$ and the state is $\omega \sim F(\Omega)$. In this case, the agent chooses an action $a \in A$ and observes a binary outcome that is either a success or a failure. Denote the outcome by $o \in \{s, f\}$. The probability of observing a success is increasing in θ and in ω. Whenever the agent observes a success, he gets a payoff $v > 0$ and whenever the outcome is a failure, the payoff is 0. In addition, the probability of success is such that for each state, there is a unique optimal action that maximizes the agent's expected payoff. Therefore, the probability of success can be seen as a monotone transformation of the utility from Framework 1.

I focus on two nested theories that have been widely studied within this framework: Full Bayesian updating and self-serving attribution bias. I explain each of these classical models of belief updating in what follows

2.2.1 The Bayesian

A Bayesian agent simultaneously updates their beliefs about θ and ω by using Bayes' rule. The posterior at period t+1 about θ after observing outcom o is given by:

$$p_{t+1}(\theta, \omega | h^t) = \frac{p[o|\theta, \omega] p_t(\theta, \omega)}{\sum_{(\theta', \omega')} p[o|\theta', \omega'] p_t(\theta', \omega')}$$

Where p_t is the belief at the start of period $t+1$ and includes all the information gathered so far.

Bayesian agents choose the effort level that maximizes their expected flow payoff by

taking expectations over their prior beliefs about θ and ω. Since agents are myopic, even though all the parameters could be identified with enough variation in choices, there will be instances in which the Bayesian will not learn even with infinite amounts of data. This happens because, by being myopic, they do not internalize the tradeoff between flow payoff and learning. This can result in too little experimentation to learn their true type. An alternative to this approach is given by Hestermann and Yaouanq [2021] and is discussed in the concluding remarks. Regardless, this approach is useful as a benchamrk for the other theories discussed in this paper.

Also notice that if a fully Bayesian agent has a dogmatic prior, they will never update their beliefs about the parameter that they are dogmatic about. This will imply that they are prone to the same types of errors as the dogmatic modeler in framework 1.

2.2.2 The Self-Serving Updater

A Self-Serving Bayesian is an agent who uses a biased version of Bayes rule to update their beliefs. They will update their beliefs about the state ω and his type θ simultaneously by over-attributing successes to a high value of θ and under-estimating the role of higher ω. Similarly, he will attribute failure to a low state to a greater degree than an unbiased agent would. To model the self-serving attribution bias, I take the approach of Benjamin [2019], where the posterior odds are given by:

$$p_{t+1}(\theta, \omega | h^t) = \frac{p[h^t|\theta, \omega]^{c(\theta,\omega,o_t)} p_t(\theta, \omega)}{\sum_{(\theta',\omega')} p[h^t|\theta', \omega']^{c(\theta,\omega,o_t)} p_t(\theta', \omega')}$$

with $c(\theta_H, \omega, o) < c(\theta_M, \omega, o) < c(\theta_L, \omega, o) \leq 1$ and $c(\theta, \omega_L, o) < c(\theta, \omega_M, o) < c(\theta, \omega_H, o) \leq 1$.

This formulation of the bias means that the agent will over-attribute successes to a high type and under-attribute them to a low type Whereas they will over-attribute failures to a low state and under-attribute them to a high state. The bias is introduced by distorting the perceived likelihood that the signal was generated by either high values of θ when it is a

success, or by low values of ω when it is a failure. The parameter c determines the degree of bias and imposes enough restrictions to ensure that the model is still falsifiable.

3 A Unifying Example

In order to compare the predictions of the theories discussed above, I develop a unifying example that allows me to isolate the forces behind each of the theories. The example is a modification of the one in Heidhues et al. [2018] and is adapted to implementable in the lab.

The agent can be of one of 3 types: $\theta \in \{\theta_L, \theta_M, \theta_H\}$ with $\theta_H > \theta_M > \theta_L$. They face an unknown exogenous success rate $\omega \in \{\omega_L, \omega_M, \omega_H\}$ with $\omega_H > \omega_M > \omega_L$. Each of the values of ω is realized with equal probability. The agent knows the distribution of ω but not its realized value.

Denote the true type by θ^* and the true state by ω^*. The agent holds some prior belief about θ[16] and chooses a binary gamble $ein\{e_L, e_M, e_H\}$. The agent observes whether the gamble is a success or a failure and gets a payoff of 1 and if it is a success; they get 0 otherwise.

The probability of success is increasing in both θ and ω and is fully described by the Table 1

	ω_H	ω_M	ω_L
e_H	50	20	2
e_M	45	30	7
e_L	40	25	20

θ_L

	ω_H	ω_M	ω_L
e_H	80	50	5
e_M	69	65	30
e_L	65	45	40

θ_M

	ω_H	ω_M	ω_L
e_H	98	65	25
e_M	80	69	35
e_L	75	55	45

θ_H

Table 1: Probability of success for each type, gamble and effort level

Conditional on a type, the agent's flow payoff is maximized by choosing the gamble that matches the state. For example, if the value of ω is ω_H, the agent's flow payoff is maximized by choosing e_H and if the state is ω_L the flow payoff is maximized by choosing gamble e_L, regardless of the value of θ. The agent myopically chooses gambles every period to maximize

[16] which is potentially misspecified as in the dogmatic and switcher cases discussed above

the flow payoff for $T < \infty$ periods.

After observing the outcome of each gamble, the agent updates their beliefs using some procedure and moves on to the next period.

Notice that both θ and ω can be identified from the outcomes if enough variation in the effort choices exists. This can be seen by confirming that there is no pair of θ and ω such that the probability of success is the same for all effort choices. Thus, by changing the effort choice, the agent can learn both their type and the state if they observe enough outcomes.

In this example, for an agent with a dogmatic belief about their type, a self-defeating equilibrium is one in which the agent chooses an effort level that, under the true θ, yields a frequency of success that is consistent with the agent's misspecified belief. That is $P[\text{sucess}|\theta^*, e^*] = P[\text{sucess}|\hat{\theta}, e^*]$ where e^* is the agent's myopic optimal choice.

In the data-generating process described above, there are five such equilibria. For example, if the agent is of type θ_M but mistakenly believes that he is of type $\hat{\theta} = \theta_H$ and the and $\omega^* = \omega_M$, when the effort chosen is e_L, the agent will observe a success with 45% chance. Because the agent dogmatically believes that their type is high, they will erroneously conclude that the rate is ω_L. Under this belief, the optimal action is e_L which will continue to generate successes with 45 probability, further reinforcing the incorrect belief. By doing so, the agent forgoes the payoff from gamble e_M which would yield a success with 65% chance.

By including self-confirming equilibria, the example captures the forces from each of the updating mechanisms discussed in the previous section and allows for the comparison of the main forces behind the theories. For realizations of (θ, ω) for which there are self-confirming equilibria, the dogmatic agent will fall into the trap whereas the switcher will be able to escape it. Similarly, an agent with self-attribution bias will update their beliefs differently from an unbiased Bayesian, leading them to choose different gambles. I exploit such cases in order to test which model is a better fit for how subjects behave in a laboratory experiment.[17]

[17]because the setting does not match that of @Heidhues2018, there will be situations for which the theory does not provide a prediction. If such cases arise in the lab, they will not be used for the analysis. However, whether a misspecified belief persists or not for the switcher, depends highly on the realized history of signals that he gets.

In what follows I explain the details of how this example was implemented in the lab.

4 Experimental Design

I recruited 78 undergraduate subjects from the CESS lab at NYU who participated in an in-person experiment. Sessions lasted approximately 45 minutes and subjects earned an average payment of $22.6. The experiment was programmed using oTree [Chen et al., 2016].

The experiment consisted of 2 treatments: the *ego-relevant* condition and the *stereotype* condition. Subjects participated in only one of the treatments. All subjects within a session participated in the same treatment and the first 4 treatments were assigned the ego-relevant condition; the rest were assigned to the stereotype treatment. The tasks were identical across treatments, except for parameter θ. In the ego-relevant condition θ is the subject's own performance in a quiz, while in the stereotype condition, it is the performance of a randomly selected subject from another session.

The experiment had 3 parts. In Part 1 subjects had 2 minutes to answer as many multiple-choice questions as they could from a 20-question quiz. They did this for quizzes on 6 different topics. The topics were: Math, Verbal Reasoning, Pop-culture and Art, Science and Technology, US Geography, and Sports and Video Games. In this part, they did not know how many questions were available and they were given no feedback.

After taking all 6 quizzes, they proceeded to part 2 where they were asked to guess their score on each of them. In the stereotype treatment they were additionally asked to guess the score of a randomly drawn participant from a previous session. All they knew about the other participant was their gender identity and whether they were US nationals or not. For each guess, they had three score options: Low-Score (5 or fewer correct answers), Mid-Score (between 6 and 15 correct answers), High-Score (16 or more). Each of the score categories corresponded to θ_L, θ_M, and θ_H respectively. They were also asked to say how confident they felt about their choices. They had 4 possible answers: "it was a random guess", "there is

another equally likely score", "I am pretty sure", "I am completely sure". These 4 answers are mapped to priors that place probabilities .33, .50, .75, and 1 to the chosen type. The remaining probability is split equally among the other two types. Questions in Part 2 were not incentivized, but subjects were told that providing an accurate answer would increase their chances of earning more money in the last part of the experiment.

The purpose of Part 2 is to classify subjects into overconfident, underconfident and correctly specified. If a subject guesses their score to be in a higher (lower) category than their true score, they are overconfident (underconfident); if they guess their score to be in the same category as their true score, they are correctly specified. This classification is done for each of the 6 topics separately.

Finally, in Part 3, subjects completed a belief updating task for each of the quizzes. Before starting the task they were reminded of their guess for the score. In the ego-relevant treatment, they were reminded of their guess about themselves and in the stereotype treatment they were reminded of their guess about the other participant. In the stereotype treatment, they were also reminded of the characteristics of the other participant.

For one topic at a time and in random order, they were presented with the three gambles from the example above and were asked to choose one of them. The probability of success was determined by their own score in the ego-relevant condition, and by the score of the other participant in the stereotype condition. Subjects had access to the three probability tables in the printout of the instructions at all times and the meaning of each cell was explained in detail.

In the interphase, they had to choose which of the 3 tables they wanted to see before entering their choice in it. This was done as an alternative to a belief elicitation in each round. I take their choice of table to be indicative of their beliefs about the underlying type. I chose not to elicit the beliefs at each round to stay true to the forces in framework 1.

Once they have entered their choice, they observe a sample of 10 outcomes from the gamble they chose. After observing the outcomes, they returned to the choice screen and

entered a new choice. In the choice screen subjects had access to the entire history of gambles and outcomes for that task as well as a summary of the outcomes so far. Once they entered 11 gambles (and observed 110 outcomes), they moved on to the next topic and repeated the same procedure. They all did this for all 6 topics.

At the end of the experiment, one of the 6 topics was randomly selected to determine the payment. They earned \$0.20 for each correct answer in the quiz, and for each success in the task in Part 3 for the selected topic.

Randomness is controlled throughout the experiment and sessions by setting a seed at the beginning of the first session. The seed was drawn at random and remained fixed for all sessions.[18] By doing this I ensure that any two subjects who have the same type and face the same exogenous rate will observe the same outcomes. This feature allows me to identify differences across subjects in updating procedures since if they use the same updating procedure, they should be choosing the same gambles.

5 Predictions

In this section, I outline the behavior that is predicted by each of the theories discussed above.

5.1 Dogmatic Modeler

Since the domain of the problem is discrete and finite in the example, the predictions of the original theory apply only to the combinations of parameters and initial beliefs for which there is a self-defeating equilibrium. In the example, there are 5 such combinations. For each of them, the dogmatic modeler predicts that the agent will fall into the self-defeating equilibrium and will be able to sustain the misspecified belief forever.

[18]The seed that was drawn at the beginning of session 1 was 3452. The same seed was used for all sessions. It is used both for drawing ω for each of the tasks in the experiment, as well as for drawing the outcomes from the gambles.

Table 2 describes the 5 self-defeating equilibria and the effort choices that sustain them. The first columns describe the combination of parameters and initial beliefs. The last column describes the effort choice that the agent will make in the long run. It is only for these combinations of parameter values and beliefs that the dogmatic model makes predictions within the Unifying example.

Prediction 1D: *Whenever an agent is of a type θ^* but mistakenly believes that they are of a type $\hat{\theta}$, and $(\theta^*, \omega^*, \hat{\theta})$ are such that there is s self-defeating equilibrium, the agent will fall into the trap and choose the effort level that sustains the misspecified belief forever.*

The model does not make predictions about what happens in cases where there is no stable belief. I assume that because there is no stable belief, the agent will eventually have to use some procedure to revise their belief about θ. In such cases, I aim to determine which of the alternative explanations provided by the other theories is a better fit for the data.

True Type (θ^*)	**True State (ω^*)**	**Believed Type ($\hat{\theta}$)**	**Believed state ($\hat{\omega}$)**	**Effort**
θ_L	ω_H	θ_M	ω_L	e_L
θ_M	ω_L	θ_L	ω_M	e_M
θ_M	ω_M	θ_H	ω_L	e_L
θ_M	ω_M	θ_L	ω_H	e_H
θ_M	ω_H	θ_H	ω_M	e_M

Table 2: Stable beliefs and the effort choices that support them for the unifying example

Although the dogmatic model does not apply to all possible parametrizations and beliefs, whether subjects fall into the self-defeating equilibria or not is still informative of the updating procedure that they are using. Understanding if the presence of traps is a key feature preventing subjects from learning the optimal action is important for understanding the prevalence of overconfidence. Similarly, gaining insight into what happens when there are no traps is important for understanding what are the other reasons why overconfidence might arise and prevail.

5.2 Switcher

Since the switcher starts as the same dogmatic agent, the initial behavior of both types of agents is identical, however, because the switcher is keeping track of the likelihood ratio, they will be able to escape the self-defeating equilibrium if the evidence is convincing enough. Therefore there is a positive probability that the switcher will adjust their initially misspecified belief about θ and learn the true state.

Prediction 1S: With positive probability, the switcher will escape the self-defeating equilibrium and learn the true state.

One caveat is that when the switcher and the dogmatic agent both start with a correctly specified belief, neither of them will fall into the self-defeating equilibria and thus will look identical even in the long run. This means that in order to distinguish between the two theories, I need to look at cases where the agent starts with a misspecified belief.

The probability that the switcher will escape the self-defeating equilibrium depends on the prior belief about ω. If the prior is sufficiently tight around the self-defeating equilibrium, the likelihood ratio will not grow fast enough, however, in the example above, the prior is uniform over the states and therefore, the likelihood ratio is more likely to grow fast enough.

5.3 Self-Attribution Bias

The key feature of Self-attribution bias is the asymmetric treatment of good news and bad news. In particular, the agent will over-attribute successes to a high type and under-attribute them to a low type. Similarly, they will over-attribute failures to a low state and under-attribute them to a high state. This implies that, after observing a failure, the agent will adjust their effort downwards by more than what an unbiased Bayesian would have done. In contrast, after observing a success, the agent will adjust their effort upwards by less than a Bayesian would have done and if the bias is large enough, it could be that the agent will not adjust their effort upwards at all, or even decrease their effort in response to a success.

This is in stark contrast with what a Dogmatic modeler would do. A dogmatic modeler

always attributes any variation in the outcome to the state and never updates their beliefs about θ. Therefore, they will always increase their effort choices after a surprising success and decrease it after a failure.

On the other hand, the biased behavior can be in line with the behavior of a switcher. In particular, if the agent starts with a misspecified belief on $\hat{\theta}$, and is willing to switch to a belief with $\theta' > \hat{\theta}$, whenever the paradigm shift happens, it will likely be in response to a surprising streak of successes. In this case, when they adjust their belief about θ, they will also adjust their effort choice. Since they were initially underconfident, they had been choosing an effort that was too high relative to the true state, and therefore, the effort is likely to fall in response to a surprising streak of success.

Prediction 1A: *After observing a failure, the agent will adjust their effort downwards by more than a Bayesian would have done. After observing a success, the agent will adjust their effort upwards by less than a Bayesian would have done. If the bias is large enough, the agent might not adjust their effort upwards at all, or even decrease their effort in response to a success.*

Prediction 1A helps distinguish between the self-attribution bias and the dogmatic modeler. However, it does not help distinguish between the self-attribution bias and the switcher. In order to do so, I need to look at cases where the agent starts with a correctly specified belief. In these cases, the switcher is very unlikely to switch to a different belief and therefore, the behavior of the switcher will mostly change in the same direction as the information he receives. [19] Instead, the self-attribution bias will lead to a different behavior. In particular, even for agents who have a belief that is correctly specified and place a lot of weight on the correct values of θ, the bias will cause these agents to become overconfident after a sequence of successes. This force is not present in the switcher or the dogmatic and it is the key feature that allows me to identify a biased updating procedure.

Prediction 2A: *After observing a streak of successes, the agent will update their belief*

[19] If they get a streak of failures they are likely to adjust their effort downwards and if they get a streak of successes they are likely to adjust their effort upwards.

about θ upwards. This will lead to the possibility of observing subjects who start correctly specified and become overconfident.

5.4 The Bayesian

The Bayesian serves as the closest benchmark to a fully rational agent. However, as discussed above, they are still vulnerable to falling into self-defeating equilibria, in which case, their long-run behavior will be identical to that of the dogmatic modeler. The main distinguishing feature is that when they do fall into a self-defeating equilibrium, there is no bias in the direction of the misspecification that they end up being trapped in. In contrast, the dogmatic modeler will always be trapped in a misspecified belief that they started with and the self-attribution updater will always become overconfident.

Prediction 1B: *When an agent starts with a diffused prior and falls into a self-defeating equilibrium, they will be equally likely to fall into the underconfident trap as the overconfident trap.*

Prediction 2B: *Bayesian agents always update their effort choices in the same direction as the information they receive. That is, if they see successes, they would only increase their effort; and if they see failures they would only decrease their effort.*

6 Stereotypes

So far I have focused only on the case of overconfidence about an ego-relevant parameter. However, each of the models can also be used to explain the prevalence of stereotypes. An agent can have a dogmatic belief about the ability of a particular group of people and either under or overestimate the associated parameter. Similarly, they can be willing to switch between two dogmatic beliefs about their ability as a switcher would. Finally, they can have a biased updating procedure in which they treat good news and bad news asymmetrically.

As such, all the predictions discussed above apply to the case of Stereotypes as well.

My analysis of stereotypes will focus only on the degree to which ego-relevance of the type affects the updating procedure and not on the motivations behind the bias. I am particularly interested in whether the explanatory power of the models differs with the degree of ego-relevance of the type.

7 Results

7.1 Initial Beliefs and misspecifications

As mentioned in the predictions section, the dogmatic modeler and the switcher are only distinguishable when the agent starts with a misspecified belief. Figure 1 shows the distribution of misspecifications by treatment. The histogram considers the difference between the subject's true score in the quiz and their guess about their score. If their guess is in a higher category than their true score, they are overconfident; if it is in a lower category, they are underconfident; and if it is in the same category, they are correctly specified. Overall, 0.43 of the guesses made in the questionnaire were misspecified. The rest of the guesses coincided with the true score. And the distribution of misspecifications is similar across treatments.

Although the overall distribution is similar for both treatments, the misspecifications arise for different combinations of characteristics in each treatment. Figure 2 shows a heatmap of the misspecifications that arose in each treatment. In the stereotype treatment, subjects are most underconfident about the performance of other non-American females in Sports and Video Games. In contrast, they are most overconfident about the performance of American men in Verbal Reasoning. In the ego-relevant treatment, subjects are most underconfident about their own performance in pop culture and art, while most overconfident about their performance in Verbal Reasoning.

Having widespread misspecifications in both treatments is important for the analysis because it allows me to distinguish between the dogmatic modeler and the switcher. If there were no misspecifications, the two models would be indistinguishable.

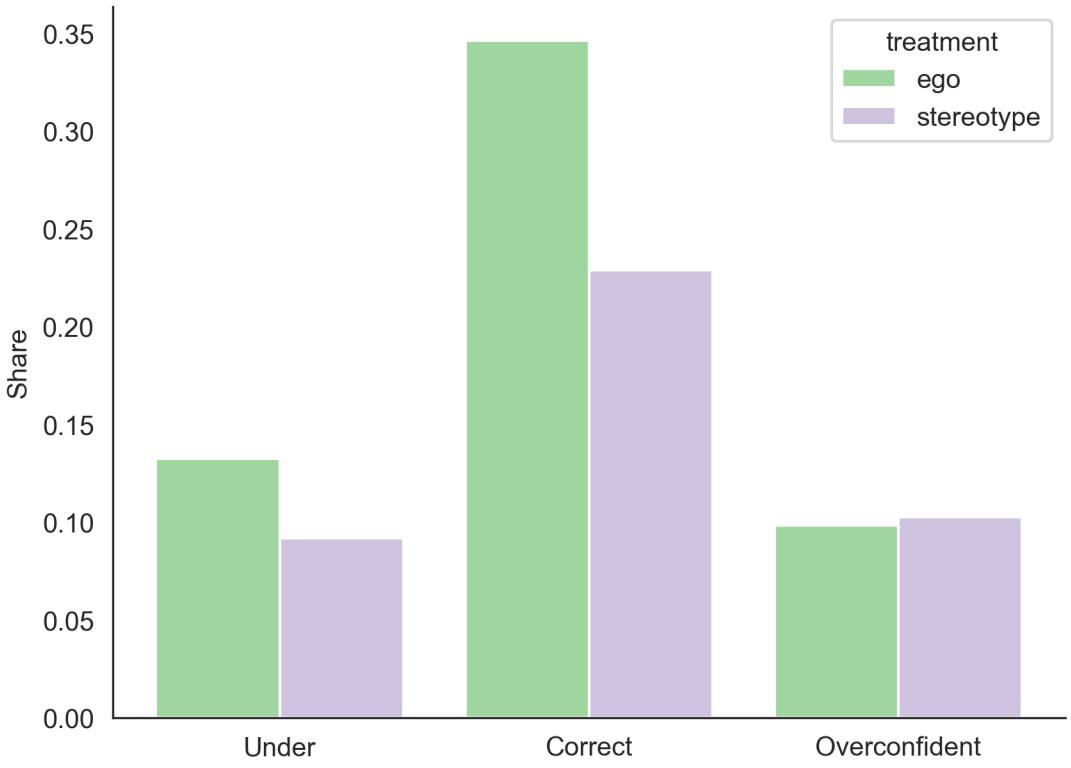

Figure 1: Initial misspecifications by treatment

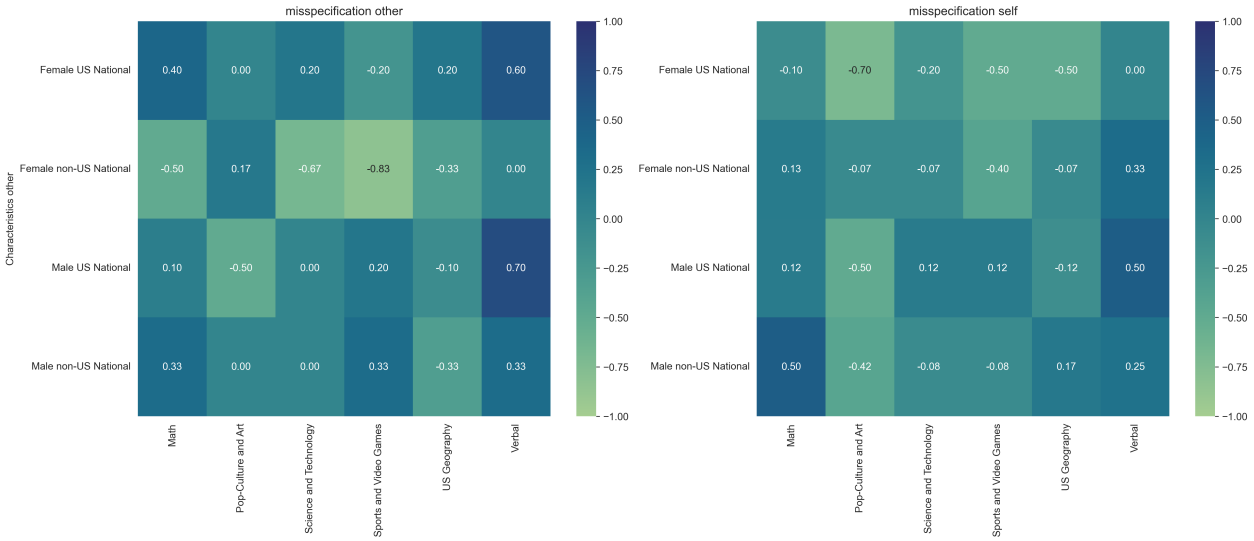

Figure 2: Initial misspecifications by treatment and characteristics

7.2 Learning

In this section, I analyze the learning behavior of subjects in the experiment. There are two parameters that subjects can be learning about: the exogenous parameter ω and the type θ. Their belief about the exogenous parameter is tracked by their choice of effort, while their belief about the type is tracked by their choice of a matrix in which to enter their effort.

7.2.1 Learning about the state

In order to analyze the learning about the state, I look at the share of optimal choices that are made at each round. I find that although subjects seem to be improving in their choices overall, the last choice coincided with the true state only in 52% of the choices in the last round of each task. This is statistically greater than the initial share of optimal choices ($p < 0.01$). However, it is still far from complete learning. It is also important to note that learning is similar across treatments. The share of choices that were consistent with the true state for each round is reported in panel A of Figure 3.

A closer look at whether people learned or not reveals that there is a good amount of heterogeneity in the sample. Panel B of Figure 3 shows the share of optimal choices by round for subjects that chose an effort that matched the state in 3 out of the last 4 rounds. It also shows the share of optimal choices for subjects who chose an effort that matched the state in fewer than 3 out of the last 4 rounds. I label the former as learners and the latter as non-learners, with learners making up 39% of the sample.

In what follows I will try to disentangle the reasons for the lack of learning. I will argue that it is not due to the presence of traps, but rather due to biased updating. According to the theories, the main reason why subjects might not learn is that they have an incorrect belief about the type or they develop an incorrect belief about the type.

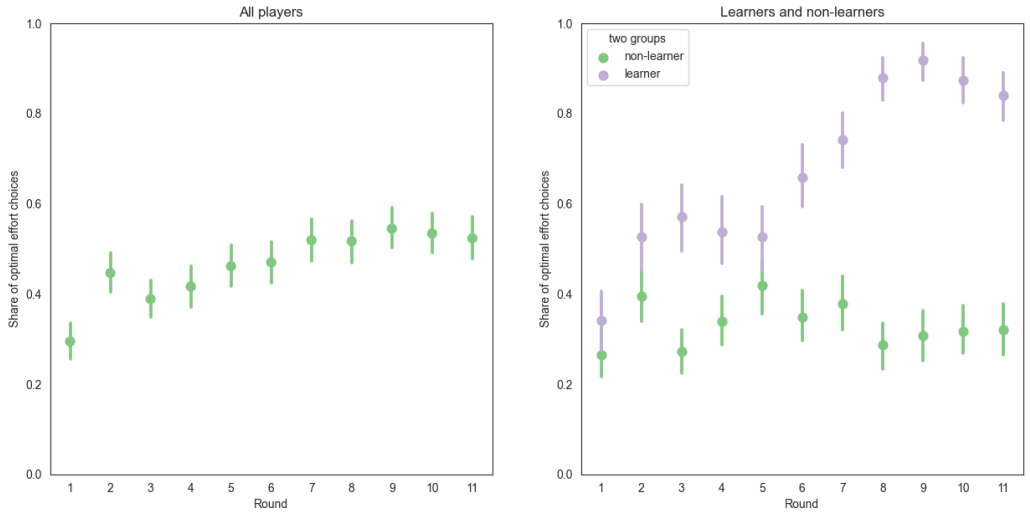

Figure 3: Share of optimal choices by round for subjects who learned (their effort in at least 3 out of tha last 4 choices matched the state), and subjects who did not learn

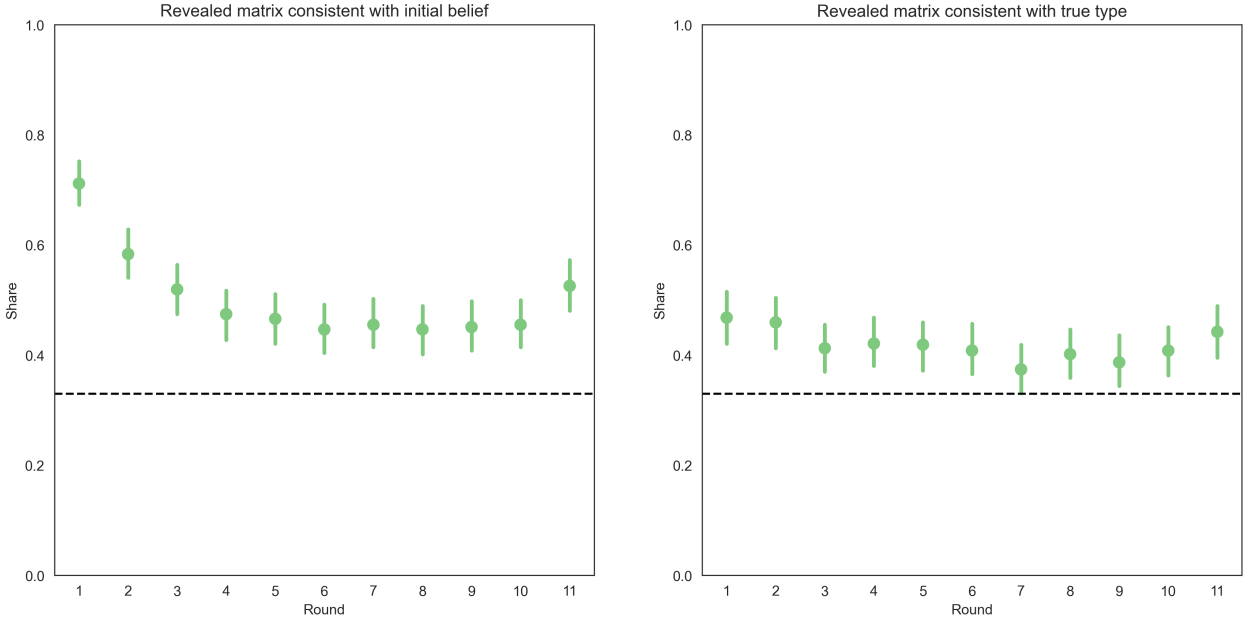

Figure 4: Matrix choices by round

7.2.2 Learning about the type

Since the elicitation of the belief about the type was not incentivized and not elicited in a standard way, I first need to confirm that subjects were not just randomly choosing a matrix in which to enter their effort. The left panel of Figure 4 shows the share of subjects who chose a matrix consistent with their initial reported belief. In round 1, 71% of the subjects chose a matrix that was consistent with their initial belief. This indicates that subjects were not just randomly choosing a matrix in which to enter their effort. From round 2 onwards, the share steadily declines, but still not as far as to indicate a random choice of matrices. This is consistent with the subjects moving away from their starting belief through some updating procedure.

On the right panel of Figure 4 I show the share of subjects who chose a matrix that is consistent with their true type. Unlike the left panel, there is no clear trend, which indicates that although they are moving away from their initial belief, they are not moving towards their true type, which means that overall, misspecification is not decreasing. A closer look at the data reveals a good amount of heterogeneity in the underlying behavior.

Figure 5 shows the transition matrix for subjects who started in each of the 3 possible starting specifications and the specifications that they ended up in at the end of the updating task. The two things to note are that the initial belief is the most likely end belief. This is consistent with some degree of stickiness of the misspecifications as the switcher and dogmatic models would predict. However, the data also presents a lot of subjects who started with a correct belief of the type and ended up overestimating it. This is consistent only with the self-attribution bias. Lastly, the subjects who initially overestimate the score, are the least likely to learn their true type.

The transition matrix points towards the presence of self-attribution bias in the data. However, it is still possible that there is heterogeneity in the updating procedure across subjects. In the next section I will look at the role of self-defeating equilibria in preventing learning. I find that the presence of self-defeating equilibria does not seem to be the main

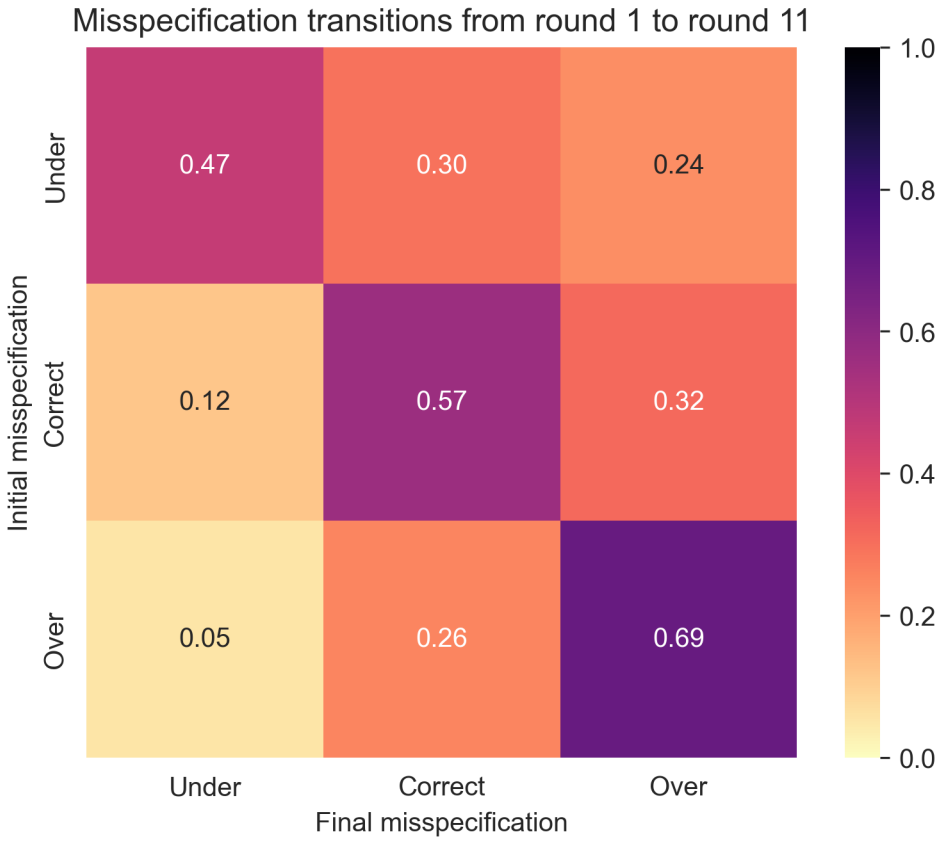

Figure 5: Transition matrix for subjects who started in each of the 3 possible starting specifications and the specification that they ended up in at the end of the updating task

deterrent to learning.

7.3 Learning traps

As described above, the presence of stable misspecified beliefs act as learning traps. If an agent falls into one of these traps, bayesian updating will lead them to choose the same effort forever. These self-defeating equilibria allow me to distinguish between the dogmatic modeler and the switcher. If the agent falls into a trap and stays there forever, they are more likely to be a dogmatic modeler. If they escape the trap, they are more likely to be a switcher or to be updating with some attribution bias.

Overall, it does not seem to be the presence of traps that is preventing learning. Panel A of Figure 6 shows that the share of subjects who learned in cases where there was a trap is larger than the share of subjects who learned in cases where there was no trap. This is true for both treatments. This indicates that even in the presence of traps, 44% of subjects learn. From the share that did not learn when there was a trap. A remaining 21% can be accounted for as being trapped.[20] While the rest of the subjects neither learned nor were trapped.

A closer look at the behavior of only the cases in which there were traps reveals that among those who did not learn, the choices that support stable beliefs were chosen much more often than among subjects who learned. Figure 6 shows the share of choices that were consistent with a stable belief for the learners and the non-learners. The difference is statistically significant ($p < 0.01$). Thus, although the presence of self-defeating equilibria is not the main deterrent to learning, it does seem to be a factor in the behavior of those who do not learn.

So far I have accounted for 16% subjects as being trapped, and 39% having learned the state correctly. From the remaining 45%, 60% were prone to traps but did not fall into them, therefore they could not have been dogmatic modelers. Alternative explanations are

[20] A subject is considered trapped when 3 out of their last 4 choices are consistent with a self-defeating equilibrium.

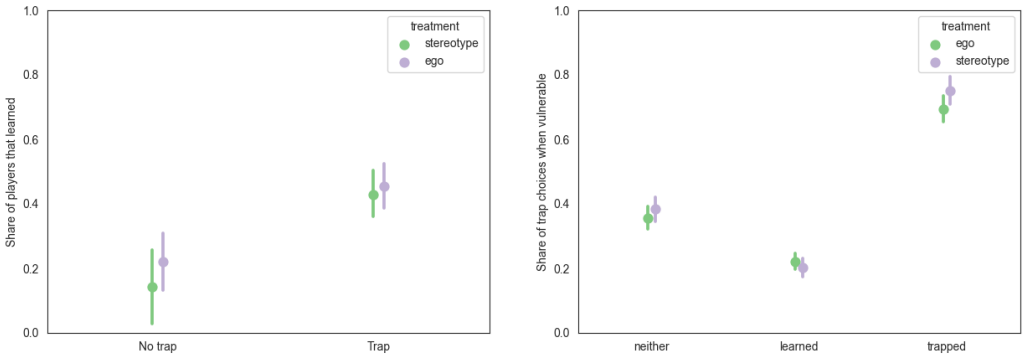

Figure 6: Panel A: Share of subjects who learned in cases where there was a trap and in cases where there was no trap. Panel B: Share of choices that were consistent with a stable belief for the learners, those who are in a trap and others

that they were switchers who considered an incorrect alternative model, or that they were updating with a bias so large that their effort choices were far from optimal. Behavior that is not consistent with either of these falls outside of the scope of the theories discussed above.

My results are consistent with the findings of Hestermann and Yaouanq [2021]. In a laboratory experiment that stays true to Framework 1, they test the predictions of the dogmatic modeler. They find that the average pattern in behavior is in the direction of the dogmatic modeler, however, subjects do not fully act as dogmatics as their actions do not go as far as to fully support the self-defeating equilibrium. A lot of the differences between the observed behavior in their experiment and the model can be attributed to subjects avoiding the self-defeating equilibria and learning the true state. In contrast to their study, I can provide more insight into the reasons why subjects avoid the self-defeating equilibria.

7.4 Reactions to good and bad news

One of the key features of the self-attribution bias is that the agent will react differently to good news and bad news. In particular, they will overreact to bad news and underreact to good news. I define good news to be any sample of outcomes for which the realized number of successes is greater than the average number of successes so far. Similarly, I define bad

news to be any sample of outcomes for which the realized number of successes is smaller than the average number of successes so far. In addition, I calculate the difference between the realized number of successes and the average number of successes so far and call it the news differential. Figure 7 shows the correlation between change in effort and the news differential, separately for good and bad news.

The overall pattern is that bad news are associated with a decrease in effort, and a more negative news differential is associated with a larger decrease in effort. On the contrary, good news are associated with a small decrease in effort. As mentioned in the predictions section, this is consistent with the self-attribution bias as well as with a change of paradigm. It is not consistent with the dogmatic modeler or the Bayesian. Because I do not directly elicit beliefs, it is not possible to distinguish between the self-attribution bias and the change of paradigm with this data alone. Nevertheless, the fact that about a third of the subjects started with a correctly specified belief and ended up overconfident is consistent only with the self-attribution bias. this points towards the fact that misattribution bias is the most likely explanation for the behavior of the subjects in the experiment.

A regression analysis of this same data reveals that indeed, the slopes are different for godd and bad news and that the difference is statistically significant.[21] The regression results are reported in Table 3. columns 2 and 3 of the table consider the model for each treatment separately. I find no significant differences in the parameters across treatment. This is consistent with the fact that the bias is not driven by the ego-relevance of θ.[22]

More convincing evidence against the Switching model is the lack of inverse or attenuated correlation between the change in effort and the news differential. In particular, if the

[21] I run a robustness check where I define good news to be a net-positive sample (half or more of the outcomes observed in the sample received in that round were successful) and bad news to be net-negative sample. I find the same pattern in this alternative regression model and the results are reported in the appendix.

[22] The ego-relevance of θ is the motivation behind the self-attribution bias, however, the bias itself is due to an incorrect updating procedure in response to good and bad news. The presence of asymmetric updating in the stereotype condition can be due to self-censoring of the subjects. In particular, subjects who initially underestimated the performance of the other participant might have realized their mistake and by overestimating it after good news. This is consistent with the data presented here as well as with the data observed in the transition matrix for stereotypes which is reported in the Appendix

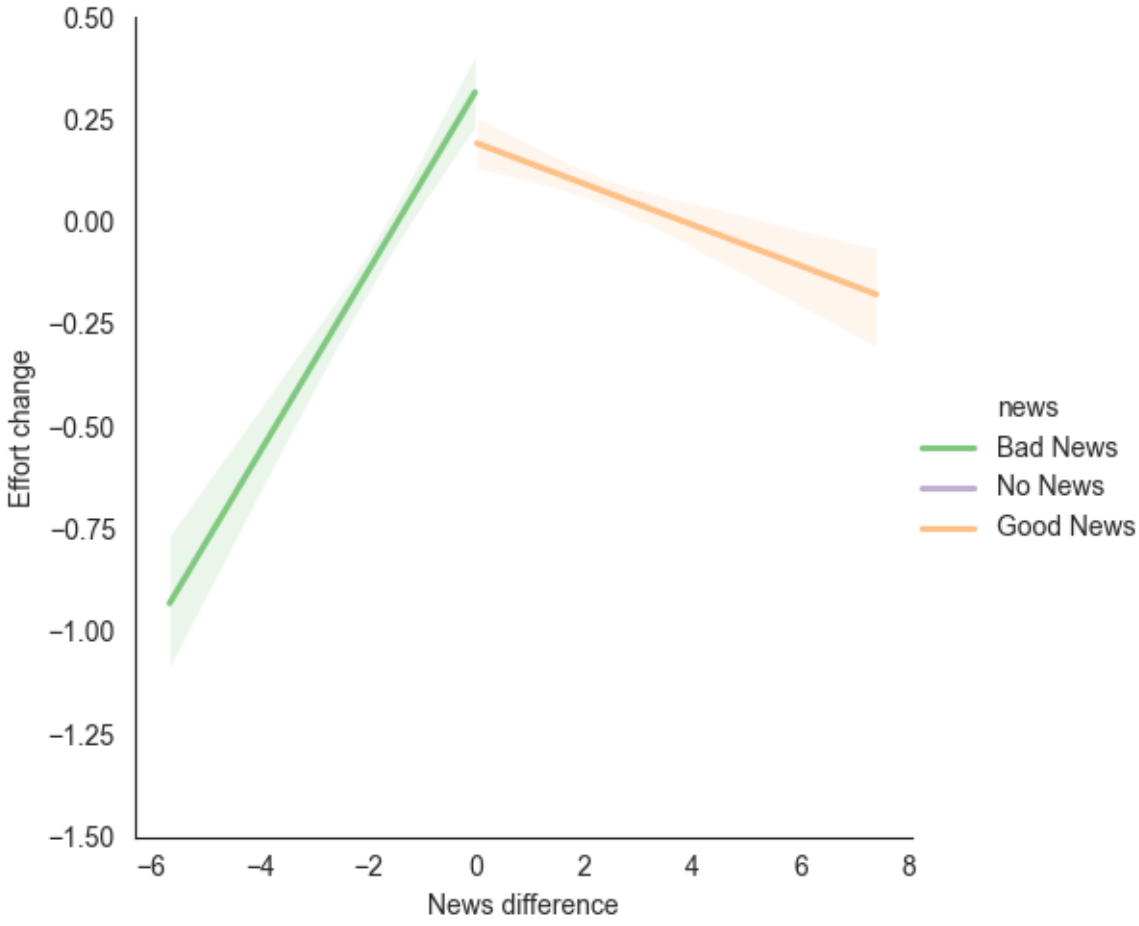

Figure 7: Change in effort by news differential

Table 3: Regression of effort change on news difference with robust standard errors

	Dependent variable:				
	Change in effort				
	All	Ego-relevant	Stereotype	Bayesian Simulation	Dogmatic Simulation
	(1)	(2)	(3)	(4)	(5)
Good news	-0.12^{**}	-0.16^{***}	-0.05	0.08	-0.08
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
News difference	0.22^{***}	0.22^{***}	0.21^{***}	0.06^{***}	0.10^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
News difference * Good news	-0.27^{***}	-0.25^{***}	-0.29^{***}	-0.04	-0.06^{***}
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Constant	0.31^{***}	0.31^{***}	0.30^{***}	-0.08^{*}	0.05
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Observations	4,680	2,700	1,980	4,680	4,680
R^2	0.04	0.04	0.04	0.05	0.06
Adjusted R^2	0.04	0.04	0.04	0.05	0.06
Note:					*p<0.1; **p<0.05; ***p<0.01

switcher were the underlying model, I would expect overconfident subjects who start with an overestimation of θ to increase their effort in response to surprising bad news. This would flatten the slope of the regression line for bad news relative to what a Bayesian would do. 3 shows that the slope for bad news for the simulation[23] of the Bayesian updates is much flatter than the slope in the data.

7.5 Stereotypes and ego-relevance

Throughout the experiment, there is little difference in the behavior of subjects across treatments. Two of the main differences are that subjects in the ego-relevant treatment are more likely to become overconfident when they start with a correct belief. And that subjects in the stereotype treatment are more likely to overreact to good news when they start by underestimating the performance of the other participant.

These two differences present as consistent with the self-attribution bias. However, the pattern in the stereotype treatment could also be due to self-censoring of the subjects. In particular, subjects who initially underestimated the performance of the other participant might have realized their mistake and over-compensated their mistake by overestimating it after good news. This is consistent with the data and offers a more intuitive explanation for why this pattern arises in the stereotype treatment.

Furthermore, I find no evidence of confirmation bias which is something that has been suggested in the literature of biased udating.[24] In particular, I find no evidence that subjects who overestimate the score of the other participant are more likely to over-attribute successes to a high type and under-attribute them to a low type. Symmetrically, I find no evidence that subjects who underestimate the score of the other participant are more likely to over-attribute failures to a low type and under-attribute them to a high type. This is consistent with what Möbius et al. [2022] find in their experiment in a less complex updating task.

[23]The details on how the data was simulated are explained in the Appendix

[24]See Benjamin [2019] for a review of the literature in this and other biases.

8 Structural estimation

I have discussed two models that are consistent with the data: the self-attribution bias and the switcher. Although the data seems to point more convincingly towards the self-attribution bias, I cannot rule out the possibility that the switcher is the underlying model. In this section, I estimate the structural parameters of the two models and try to classify subjects into each of the models to assess how much of the data is explained by each of the two likely theories.

8.1 Estimation of the switching threshold

The switching threshold is the parameter that determines the likelihood ratio at which the switcher abandons the status quo paradigm and adopts the alternative. In order to estimate this parameter, I assume that all subjects are switchers and look at the rounds in which they chose to reveal a probability matrix that was different from the one they revealed in the previous round. I then calculate the likelihood ratio for the alternative paradigm to the status quo. I average across all such likelihood ratios and use the average as an estimate of the switching threshold.

In detail, Let b_t^θ be the belief about θ in period t.[25]. And consider a subject for whom $b_{\tau-1}^\theta \neq b_\tau^\theta$, where b_t^θ. That is, in period $\tau-1$ they chose to reveal a matrix that was different from the one they revealed in period τ. The matrix revealed in period $\tau-1$ is the status quo paradigm and the matrix revealed in period τ is the alternative paradigm. The likelihood ratio for the alternative paradigm to the status quo is

$$\ell = \frac{L(\theta|b_\tau^\theta)}{L(\theta|b_{\tau-1}^\theta)} = \frac{p_\tau(o_t|b_\tau^\theta)}{p_\tau(o_t|b_{\tau-1}^\theta)}$$

where $p_\tau(o_t|b_\tau^\theta)$ is the probability of observing the signal o_t when the agent is of type b_τ^θ.

The estimate of the switching threshold, α, is the average of the likelihood ratios across

[25]in the experimental data, I use the revealed matrix as a proxy for the belief. I take the matrix revealed in the previous period to be the status-quo and the matrix revealed in period t to be the alternative

all subjects who switched in the experiment, which is given by

$$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N} \ell_{i\tau}$$

Where N is the number of switches observed in the experiment.

Because the feedback that subjects receive is the equivalent of 10 signal realizations, the estimate of the parameter is an upper bound of the true parameter. A lower bound would be given by the average likelihood ratio in the round before the switch happens. I choose to estimate the upper bound to impose a harsher test on the theory.

Because the likelihood ratio is a function of the signal, it depends on the sequence of signals that subjects observe. This means that in some cases, the likelihood ratio increases by a large amount from one period to another. Although such cases are not common[26], a few very large values can skew the average. To avoid a biased parameter, I restrict my sample to those observations for which the likelihood ratio is less than 5.[27] Under these assumptions, the estimate of the switching threshold is 1.09.

This means that the evidence supporting the alternative paradigm needs to be approximately 10% more compelling than the evidence supporting the status quo paradigm in order for the agent to switch. The estimation treatment by treatment is slightly larger for the Ego condition than for the Stereotype condition. However, the difference is not statistically significant (p-value=0.96).

I then use this estimate to simulate the behavior of the switcher to assess the fit of the model. Which is discussed in the next section.

8.2 Estimation of the self-attribution bias

I use the simulated method of moments to estimate the distortion parameters of the misattribution bias model. There is a toltal of 6 parameters in the model, however, I

[26]In the data, only 1.8observations have a likelihood ratio that exceeds 5.
[27]The full distribution of likelihood ratios is reported in the appendix

assume that the distortion is symmetric for good and bad news. Formally, I assume that $c(\theta_H, \omega, \text{good news}) = c(\theta, \omega_L, \text{bad news})$, $c(\theta_M, \omega, \text{good news}) = c(\theta, \omega_M, \text{bad news})$ and $c(\theta_L, \omega, \text{good news}) = c(\theta, \omega_H, \text{bad news})$.

To estimate them I use 4 moment conditions. The moments that I use are the regression coefficients from the reaction to good and bad news. That is: the slope and intercept of the relationship between the number of successes in the last signal sample and the change in effort as a consequence. The slope and intercept are different for good and bad news, which gives 4 moment conditions. To determine good news are samples for which there are more successes than failures. Bad news are outcome samples for which there are more failures than successes. The sample moments are reported in 4 and correspond to the following regression model:

$$\Delta effort_{it} = \beta_0 + \beta_1 signal_{it-1} + \beta_2 \mathbb{I}\{good\, news_{it-1}\} + \beta_3 I\{good\, news\} * signal_{it-1}$$

where i is an individual and task level index (each subject in each of the six tasks in the experiment represents a single observed path) and t is a round_number index that ranges from 1 to 11. The variable $signal_{it-1}$ is the number of successes observed in round $t-1$, and the change in effort is $\Delta effort_{it} = effort_{it} - effort_{it-1}$. The coefficients β_0 and β_1 are the moments that rule the interaction between bad news and effort change, while $\beta_0 + \beta_2$ and $\beta_1 + \beta_3$ are the intercept and the slope for the relation between effort change and good news.

In order to find the parameters that correspond to the observed moments I rely on simulation. Rather than theoretically deriving the moments, I simulate what a large number of subjects would have done conditional on parameter values. I do this for a large number of parameter combinations that statisfy the restrictions on them. For each set of parameter values, I compute the simulated moments. These simulations are meant to approximate the population moments conditional on parameters. I then look for the set of parameters that minimize the weighted quadratic distance between the sample moments and the population

moments. The weight I place on each of the moments is the inverse of the variance of the moment in the sample.

The model makes deterministic predictions about the choices subjects should make after each possible history of signals. However, in the data I observe significant variation. In order to account for variation in the simulation I introduce a noise parameter $\varepsilon_{it} \sim U[0,1]$ which introduces a tremble in the choice of subject i in period t. If $\varepsilon_{it} < 0.1$ the subject chooses an effort level at random. Otherwise they choose the optimal effort as predicted by the model. More details on the simulation can be found in the Appendix.

Using the estimated parameters I proceed to simulate the deterministic choice paths for each of the models.

9 Model fit

I simulate each of the models for each possible combination of type and state (θ, ω). I then compute te distance from the observed paths of each subject to each of the simulated paths that match the subjects type and state. I average across all paths for a single subject (each subject participated in 6 distinct effort tasks) and determine which model is a better explanation for that subject's behavior by finding the one that gives the minimal average distance between the observed behavior and the simulation.

In detail, let e_{ikt} be the effort chosen by subject i in task k at round t. And let $\tilde{e}_t(\theta, \omega, j)$ be the simulated choices using model j for a subject of type θ in state ω. And let j be dogmatic, switcher, attribution or Bayesian. I solve the following problem to determine the best fitting model for each subject i.

$$arg\min_{j} \sum_{k} \frac{1}{T} \sum_{t} \left[e_{ikt} - \tilde{e}_t(\theta, \omega, j) \right]$$

The histogram in Figure 8 shows the distribution of the best fitting models in the sample. The average fit of each model is reported in the Appendix.

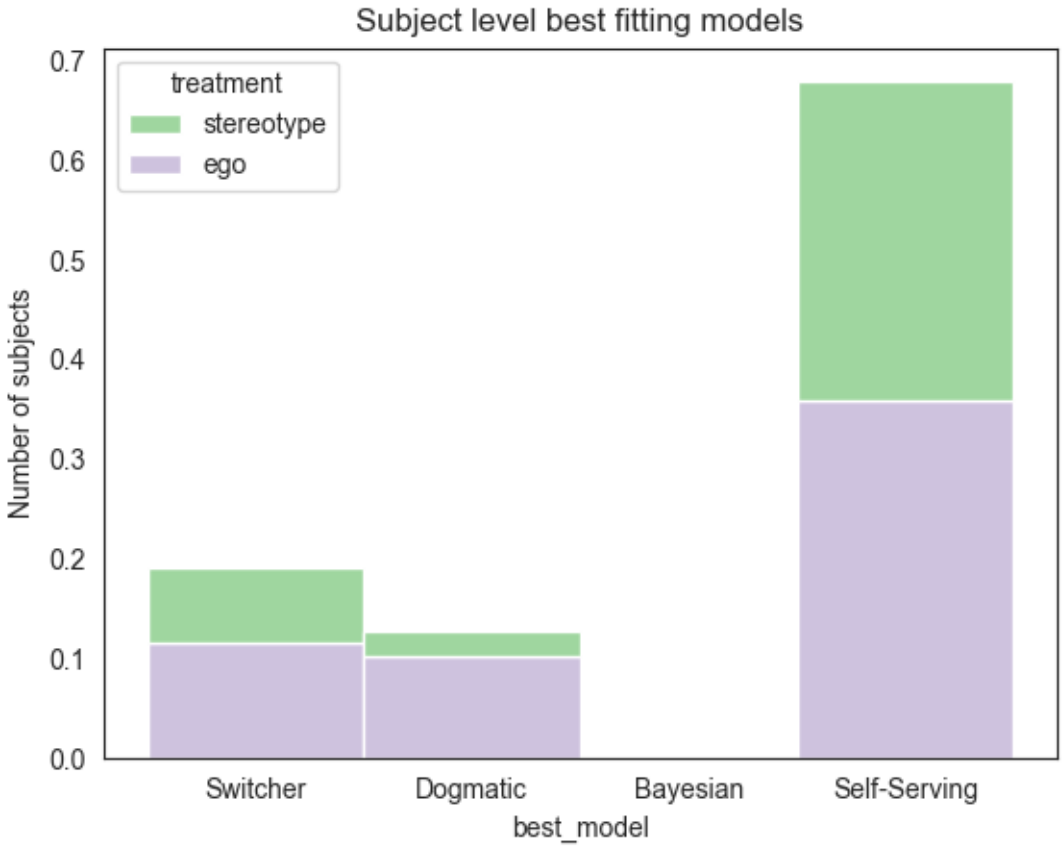

Figure 8: Distribution of the best fitting models at a subject level in each treatment

10 Conclusion

I designed a setting in which I can distinguish between the forces of dogmatism, paradigm shifts, and optimal expectations. Through a laboratory experiment, I collected data on behavior and infer beliefs from it. I find that the behavior of the subjects is consistent with the self-attribution bias. In particular, I find that subjects decrease their effort in response to bad news as well as to bad news, although the latter decrease is smaller. This is consistent with the self-attribution bias as well as with a change of paradigm. However, I also observe that about a third of the subjects started with a correctly specified belief and ended up overconfident. This is consistent only with the self-attribution bias. Therefore, I conclude that the model that best explains the data is that of optimal expectations.

I have small or no effects throughout on the ego-relevance of the type. Although the motivation behind the theories does not directly apply to the case of stereotypes, the lack of stark differences suggests that biases arise due to an incorrect updating procedure and not necessarily due to the self-serving motivation behind the bias. The lack of differences in the behavior of the subjects across treatments could also be due to self-censoring of the subjects. In particular, subjects might be initially reluctant to express their true belief about a particular group of people. In addition, as they learn and realize that their beliefs might be incorrect, they might overreact to signals to try to compensate for the initial bias. Such an updating procedure would be consistent with the self-attribution bias for the case of stereotypes. I remain agnostic about whether that is truly the case or whether the lack of differences is due to other factors.

Another bias that has been studied and suggested in the literature is the confirmation bias. In particular, it has been suggested that subjects who are overestimate θ in their prior, will over-attribute successes to a high type and under-attribute them to a low type. Symmetrically, subjects who initially underestimate θ will over-attribute failures to a low type and under-attribute them to a high type. I do not find evidence of such a bias in the data. This is consistent with what Möbius et al. [2022] find in their experiment.

I further investigate possible heterogeneity in the underlying models being used by estimating the structural parameters of the models.

A appendix

A.1 Transition matrix by treatment

Figure 9 shows the transition matrix for each of the treatments separately. It shows that initial misspecifications are more sticky in the ego treatment and that subjects who underestimate the score of the other participant in the stereotype treatment are more likely to end up overestimating it than in the ego treatment. This may signal some sort of self-censoring or overreaction in the stereotype treatment.

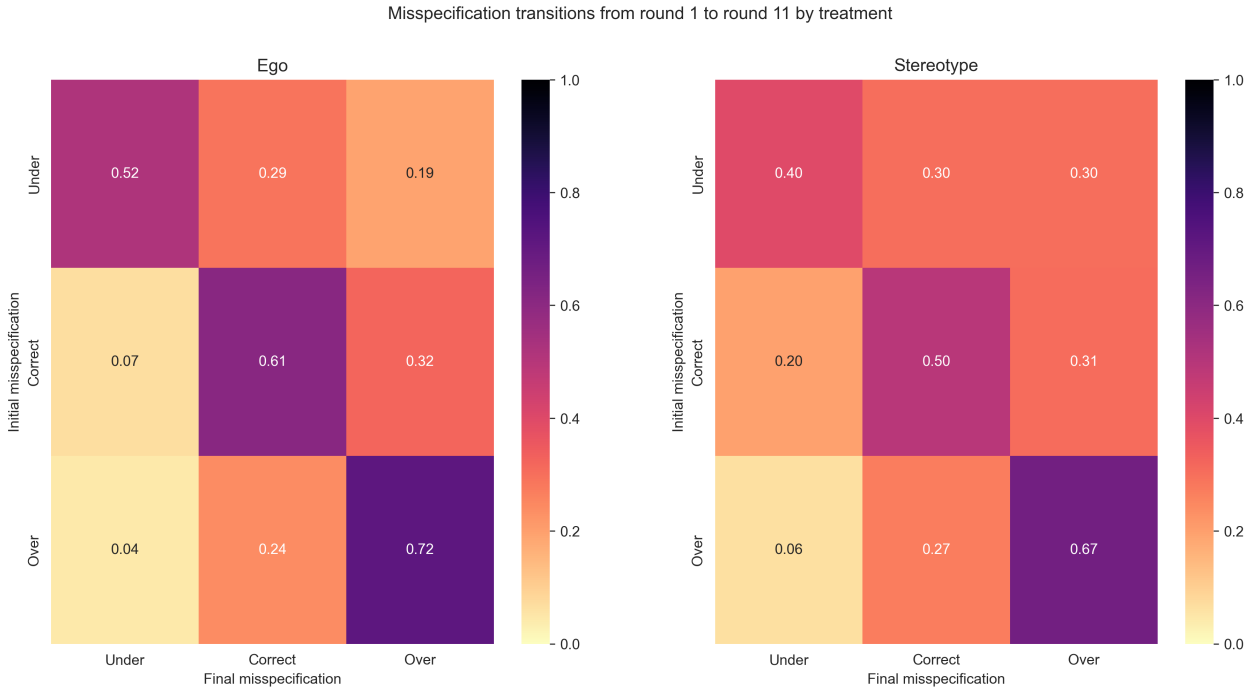

Figure 9: Transition matrix for subjects who started in each of the 3 possible starting specifications and the specification that they ended up in at the end of the updating task

A.2 Full distribution of likelihood ratios

The likelihood ratios at rounds in which I observe a change in the belief about θ presents a long right-tail. In particular, 1.8% of the observations have a likelihood ratio that exceeds 5 which would mean that the agent requires evidence that is 5 times more compelling in favor

of the alternative paradigm in order to switch. Figure 10 shows the distribution of likelihood ratios for all observations.

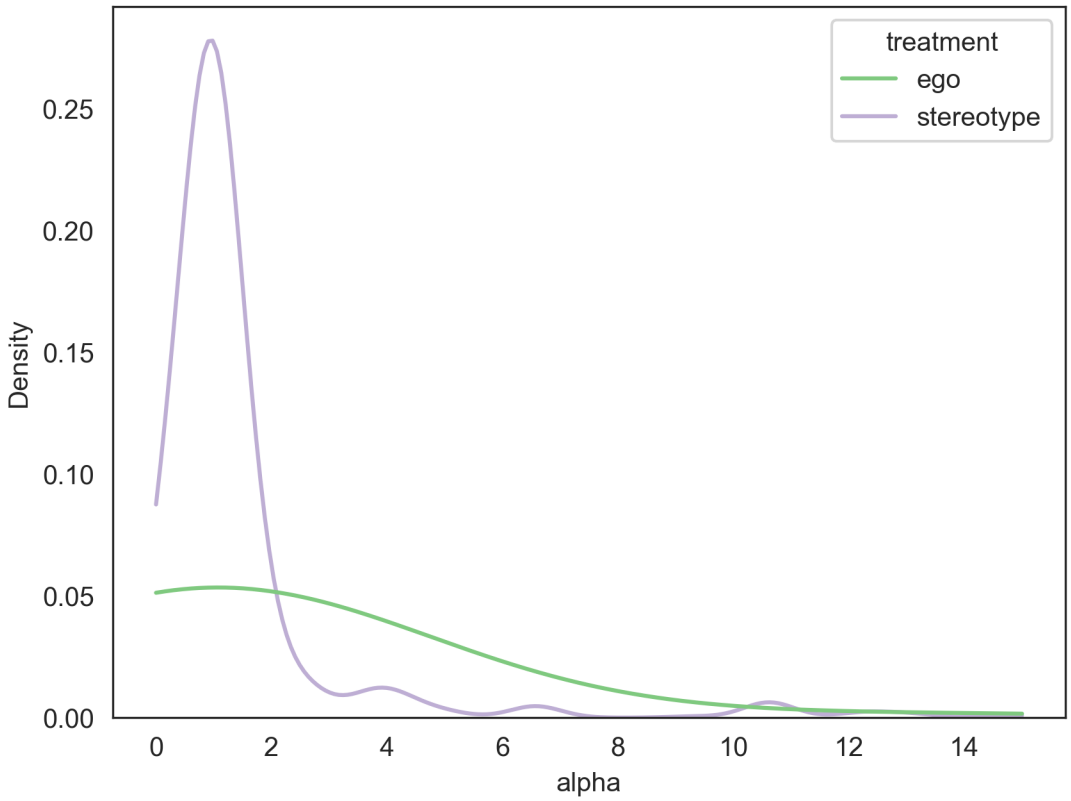

Figure 10: Distribution of likelihood ratios for all observed belief changes

A.3 Simulated Method of Moments

I simulate the model on a grid with 3 parameters: c_1, c_2 and c_3 such that $c_1 < c_2 < c_3$. All parameters on the grid take values between 0.1 and 2 with a distance of 0.1 between them. I simulate 1000 paths of choices of length 11 for each of the possible combinations of parameters. For each set of parameters I compute the moments by estimating the regression model described in section 8.2. Represent the simulated models for parameter vector c by $\tilde{m}(c)$ and the sample moments by m. The estimates of the parameters arw given by

$$\hat{c} = arg\min_{c}(\tilde{m}(c) - m)^T W(\tilde{m}(c) - m)$$

It is also worth noting that since the randomness if fixed in the experimental data by setting a seed, I do the same in the simulation and use the same seed as in the experimental environment to account for the correlation in observed signals that exists in the sample.

A.4 Model simulation

In order to simulate the data for each of the models and asses their fit, I simulated the path that each type of agent would have taken given each possible realization of (θ, ω). I fixed the seed of the random number generator to 3452[28] and fixed the signals that would be shown in each round after each choice by drawing them before simulating. This way, the signals that the simulation uses coincide exactly with the signals that a subject would have seen in the experiment. I then use each of the algorithms defined by the models to simulate the path that the agent would have taken given the signal and given an initial belief. For the dogmatic modeler and the switcher, I use each possible dogmatic belief. For the Bayesian and the self-attribution bias, I use a uniform belief on theta.

A.5 Model fit

The average model fit for each of the simulated models is reported in Figure 11.

A.6 Reaction to good and bad news (robustness check)

I use an alternate definition for good and bad news and check whether the results are robust to this. I define a signal to be a net positive if the number of successes is equal to or greater than the number of failures. I define a signal to be a net negative if the number of failures is greater than the number of successes. I now treat net positive signals as good news and net negative signals as bad news. I then rerun the regression analysis and find that the results are robust to this alternative definition. In particular, net negative signals are associated with a stark decrease in effort for both treatments. Net positive signals are associated with

[28] this is the seed that was drawn at random for the experiment

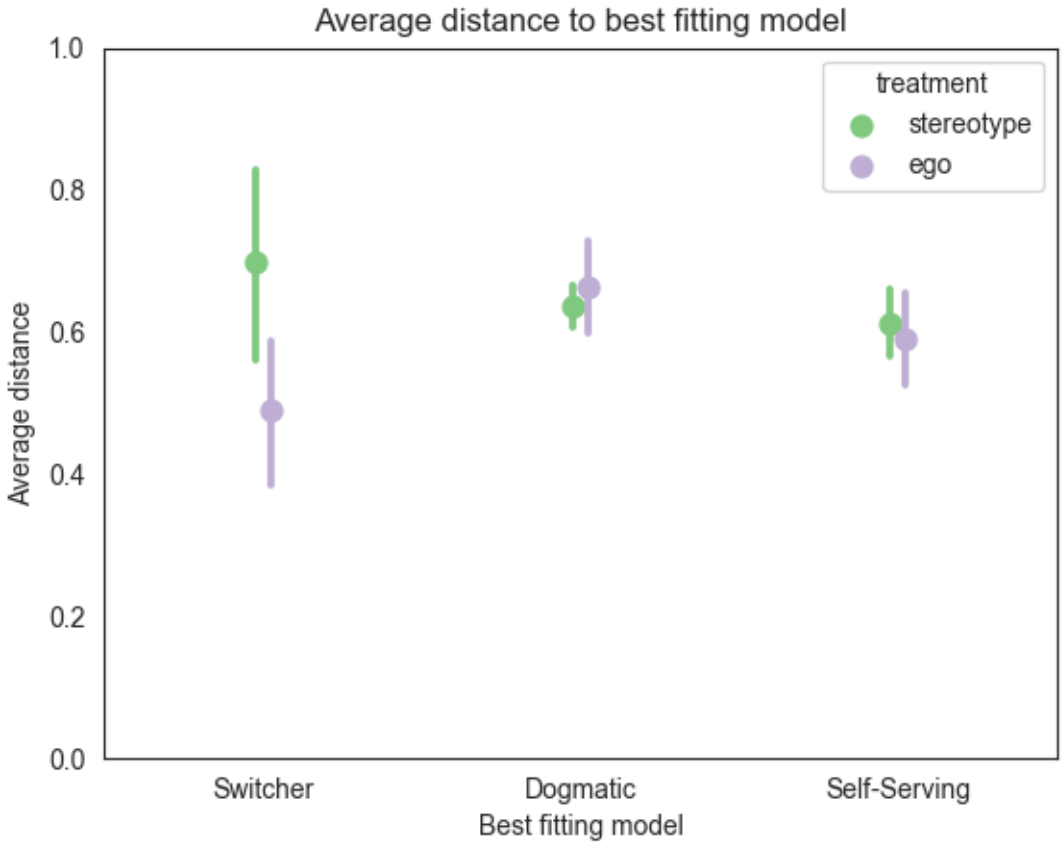

Figure 11: Average distance of subject behavior to the best fitting model

a small decrease in effort for both treatments. The results are reported in Table 4.

Table 4: Regression of effort change on news difference with robust standard errors

	Dependent variable:		
	Change in effort		
	All	Ego-relevant	Stereotype
	(1)	(2)	(3)
Net-positive signal	0.99^{***}	0.80^{***}	1.24^{***}
	(0.05)	(0.05)	(0.05)
Signal value	0.21^{***}	0.22^{***}	0.21^{***}
	(0.02)	(0.02)	(0.02)
Signal value * Net-Positive signal	-0.28^{***}	-0.26^{***}	-0.31^{***}
	(0.02)	(0.02)	(0.02)
Constant	-0.49^{***}	-0.48^{***}	-0.50^{***}
	(0.04)	(0.04)	(0.04)
Observations	4,680	2,700	1,980
R^2	0.05	0.05	0.05
Adjusted R^2	0.05	0.04	0.05
Note:			*p<0.1; **p<0.05; ***p<0.01

B Experimental Instructions

You are about to participate in an experiment on decision-making. What you earn depends partly on your decisions and partly on chance. Please turn off cell phones and similar devices now. Please do not talk or in any way try to communicate with other participants. We will start with a brief instruction period. If you have any questions during this period, raise your hand, and your question will be answered so everyone can hear.

General Instructions

The experiment is separated into two parts. You will be given instructions for each part when it is reached.

Part 1:

In part 1, you will be asked to solve a series of quizzes. In each quiz, you will answer multiple choice questions on one of 6 topics: Math, Verbal Reasoning, US Geography, Science and Technology, Pop Culture, and Sports and Video Games.

The order of the topics will be determined randomly.

You will see only one question at a time. Select an answer and then click the "Next" button to move on to the following question.

If you leave a question unanswered, it will be marked as incorrect. You will not be able to go back to that question once you click the "Next" button.

You will have 2 minutes for each quiz. Once the time runs out, your answers will be submitted automatically.

At the end of the experiment, the computer will randomly select one topic (each chosen with equal probability), and you will be paid $0.20 for each correct answer.

Your score on the quizzes will also affect how much you earn in part 2. The higher your score, the more likely you will be to make more money.

When you finish all six quizzes, there will be a short questionnaire that will not affect your payoff. Please answer all the questions as accurately as you can. You will have access to your answers from Part 1 when in Part 2, which can help you make better choices.

We will move on to Part 2 once everyone completes Part 1.

Part 2 (Ego):

The task in this part will be repeated once for each topic from Part 1. For each task, you will make 11 choices.

In this part, you will choose one of 3 gambles and see 10 outcomes for the one you choose. Each outcome will be either a success or a failure, and the probability with which each of them is a success is determined by 3 factors:

Your score on the quiz for the corresponding topic,

A randomly chosen success rate,

Your choice of a gamble.

Your score is either Low (if it is 5 or less), Mid (if it is between 6 and 15), or High (if it is 16 or more). You will not know your actual score, but you will be reminded of the guess you made in part 1.

The success rate will be chosen randomly by the computer, and it can be Rate A, B, or C. Each is drawn with an equal chance (1/3 chance each), but you will not know which one was drawn. The rate is drawn once, at the beginning of the task, and stays fixed throughout.

To maximize the chance of success, you should choose the gamble that matches the underlying success rate: - Gamble A maximizes the probability of a success when the rate is A, - Gamble B maximizes the probability of success when the rate is B and - Gamble C maximizes the probability of success when the rate is C.

After 11 gamble choices, the task will change to the next topic. This means that the probability of success for this new task will be determined by the following: - The score you received on the quiz for the corresponding topic and - a new draw of the rate (A, B, or C)

At the end of the experiment, the computer will randomly pick one of the topics, and you will be paid $0.20 for each success. (For each topic, there will be 110 outcomes: 10 outcomes for each of the 11 choices you made).

We will now go over the details of the probability of success. They are described by the tables in the back.

Three tables describe the probabilities. Each table corresponds to one of the score levels: Low, Mid, and High.

Each of the columns within the matrix corresponds to one of the success rates. You do not know which was drawn, but the gambles' outcomes can help you determine the rate.

You will choose a Row.

In order to enter your choice of a gamble, you will first need to choose which matrix you want to see. If you choose a matrix that does not correspond to your score, the probabilities in the table will not match the probabilities with which the outcomes are successful.

The outcomes will be generated using: - your actual score (table), which you do not know for certain, - the gamble (row) that you chose, and - the rate (column), which you also don't know.

Part 2 (stereotype):

The task in this part will be repeated once for each topic from Part 1. For each task, you will make 11 choices.

In this part, you will choose one of 3 gambles and see 10 outcomes for the one you choose. Each outcome will be either a success or a failure, and the probability with which each of them is a success is determined by 3 factors: - The score of another participant on the quiz for the corresponding topic, - A randomly chosen success rate, - Your choice of a gamble.

The other participant's score is either Low (if it is 5 or less), Mid (if it is between 6 and 15), or High (if it is 16 or more). You will not know the score but will be reminded of your guess in part 1.

The success rate will be chosen randomly by the computer, and it can be Rate A, B, or C. Each is drawn with an equal chance (1/3 chance each), but you will not know which one was drawn. The rate is drawn once, at the beginning of the task, and stays fixed throughout.

To maximize the chance of success, you should choose the gamble that matches the underlying success rate: - Gamble A maximizes the probability of a success when the rate is A, - Gamble B maximizes the probability of success when the rate is B and - Gamble C maximizes the probability of success when the rate is C.

After 11 gamble choices, the task will change to the next topic. This means that the probability of success for this new task will be determined by the following: - The score the other participant got on the quiz for the corresponding topic - A new draw of the rate(A, B, or C)

At the end of the experiment, the computer will randomly pick one of the topics, and you will be paid $0.20 for each success. (For each topic, there will be 110 outcomes: 10

outcomes for each of the 11 choices you made).

We will now go over the details of the probability of success. They are described by the tables in the back.

Three tables describe the probabilities. Each table corresponds to one of the score levels: Low, Mid, and High.

Each of the columns within the matrix corresponds to one of the success rates. You do not know which one was drawn, but the gambles' outcomes can help you determine the rate.

You will choose a Row.

In order to enter your choice of a gamble, you will first need to choose which matrix you want to see. If you choose a matrix that does not correspond to the score, the probabilities in the table will not match the probabilities with which the outcomes are successful.

Once you enter a choice, the 10 outcomes will be generated using the following: - the actual score (matrix), which you do not know for certain, - the gamble (row) that you chose, and - the rate (column), which you also don't know.

The full data generating process is printed on the rear side of the instructions and always available to the subjects throughout part 2. Each of the matrices is labeled with "if your/the other participant's score is..." depending on the treatment

Screen walk-through script:

This is the screen where you will enter your choices of gambles. First, you must choose the matrix that you want to see. Whatever matrix you choose to see does not change the probabilities with which the gables are a success. Then, you must choose a gamble (which corresponds to a row). The computer will draw 10 outcomes for that gamble using the probability in the chosen row and the column corresponding to the rate for this task. On the right, you can choose to see either the total count of successes and failures for each gamble or the detailed history. That is the number of successes and failures for each of your choices. The rate will be drawn again for the next task, corresponding to the next quiz topic. And you will repeat this process.

References

Cuimin Ba. Robust misspecified models and paradigm shifts. 2023.

Daniel J. Benjamin. *Errors in probabilistic reasoning and judgment biases*, pages 69–186. 2019. doi: 10.1016/bs.hesbe.2018.11.002.

Anat Bracha and Donald J. Brown. Affective decision making: A theory of optimism bias. *Games and Economic Behavior*, 75:67–80, 5 2012. ISSN 0899-8256. doi: 10.1016/J.GEB.2011.11.004.

Markus K Brunnermeier, Jonathan A Parker, Andrew Abel, Roland Bénabou, An-Drew Caplin, Larry Epstein, Ana Fernandes, Christian Gol-Lier, Lars Hansen, David Laibson, Augustin Landier, Erzo Luttmer, Sendhil Mullainathan, Filippos Papakonstantinou, Wolfgang Pesendorfer, Larry Samuelson, and Robert Shimer. Optimal expectations. *The American Economic Review*, 95:1092–1118, 2005.

Colin Camerer and Dan Lovallo. Overconfidence and excess entry: An experimental approach. *American Economic Review*, 89:306–318, 3 1999. ISSN 0002-8282. doi: 10.1257/aer.89.1.306.

Daniel L. Chen, Martin Schonger, and Chris Wickens. otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97, 3 2016. ISSN 22146350. doi: 10.1016/j.jbef.2015.12.001. URL https://linkinghub.elsevier.com/retrieve/pii/S2214635016000101.

Alexander Coutts, Leonie Gerhards, Zahra Murad, Kai Barron, Thomas Buser, Tingting Ding, Han Koh, Yves Le Yaouanq, Robin Lumsdaine, Cesar Mantilla, Luis Santos Pinto, Giorgia Romagnoli, Adam Sanjurjo, Marcello Sartarelli, Peter Schwardmann, Sebastian Schweighofer-Kodritsch, Séverine Toussaert, Joël Van Der Weele, and Georg Weizsäcker. What to blame? self-serving attribution bias with multi-dimensional uncertainty. 2020.

Ignacio Esponda and Demian Pouzo. Berk-nash equilibrium: A framework for modeling agents with misspecified models. *Econometrica*, 84:1093–1130, 2016. ISSN 0012-9682. doi: 10.3982/ecta12609.

Lorenz Götte and Marta Kozakiewicz. Experimental evidence on misguided learning *. 2022.

Paul Heidhues, Botond Kőszegi, and Philipp Strack. Unrealistic expectations and misguided learning. *Econometrica*, 86:1159–1214, 2018. ISSN 0012-9682. doi: 10.3982/ecta14084.

Nina Hestermann and Yves Le Yaouanq. Experimentation with self-serving attribution biases. *American Economic Journal: Microeconomics*, 13:198–237, 2021. ISSN 19457685. doi: 10.1257/mic.20180326.

Mitchell Hoffman and Stephen V. Burks. Worker overconfidence: Field evidence and implications for employee turnover and firm profits. *Quantitative Economics*, 11:315–348, 2020. ISSN 1759-7323. doi: 10.3982/QE834.

Harold H Kelley and John L Michela. Attribution theory and research. 1980. URL www.annualreviews.org.

Markus M. Möbius, Muriel Niederle, Paul Niehaus, and Tanya S. Rosenblat. Managing self-confidence: Theory and experimental evidence. *Management Science*, 68:7793–7817, 11 2022. ISSN 0025-1909. doi: 10.1287/mnsc.2021.4294. URL https://pubsonline.informs.org/doi/10.1287/mnsc.2021.4294.

Emily Oster, Ira Shoulson, and E. Ray Dorsey. Optimal expectations and limited medical testing: Evidence from huntington disease. *American Economic Review*, 103:804–30, 4 2013. ISSN 0002-8282. doi: 10.1257/AER.103.2.804.

Hakan Ozyilmaz. Mental models and endogenous learning. 2022. URL https://drive.google.com/file/d/1sNnaXye8p2bZ3Zzd36dYF0Dza47E26PG/view.

Joshua Schwartzstein and Adi Sunderam. Using models to persuade. *American Economic Review*, 111:276–323, 1 2021. ISSN 19447981. doi: 10.1257/aer.20191074.