Learning with misspecified models: Overconfidence and Stereotypes

Jimena Galindo

September 07, 2023

Abstract

TBW

1 Introduction

2 Related Literature

3 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 prior belief about his type is $p_0(\theta)$ and his belief about the state state $p_0(\omega)$ coincides with the true distribution f. Let the agent's type be θ^* and the realized state be ω^*.

An agent has a *misspecified* belief if the prior assigns probability zero to their true type. Furthermore, the agent is *dogmatic* if he holds a degenerate belief that places probability one on being a particular type, $\hat{\theta}$. An agent can be dogmatic and misspecified; that means that $\hat{\theta} \neq \theta^*$ and $p_0(\hat{\theta}) = 1$.

The agent chooses an action $a \in A$ and observes a noisy outcome $h \in H$. 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. $\varepsilon \sim N(0, \sigma)$ 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 is symplifies the analysis and plays a role on whether an agent who updates their beliefs using bayes rule would learn the truth or not.

A key notion in this setting is that of a self-defeating equilibrium[1]. 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. The average outcome under the true type and the true state equals the average output the agent expects under the misspecified belief. In addition, 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 and this is the sense in which the two are nested. Both theories produce different predictions with respect to the equilibrium outcome.

3.1 The Dogmatic Modeler

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

[1]This notion is an adaptation of the Berk-Nash equilibrium in @Esponda2016 to this setting with only one agent

not update his beliefs. Any discrepancies between the observed outcomes and his believed type are incorporated using the Bayes rule to update their beliefs about ω. This means the dogmatic agent will never learn their true type if they startu off as being misspecified.

Heidhues et al. [2018] show that, under certain assumptions on the per-period utility,[2] 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 whith 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 ths 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 further illustrated in Example 1.

Example 1: *Set $A = \Omega$ and $H = [0, \infty)$ and consider a student with intrinsic ability $\theta^* \geq 0$ who faces a grading procedure ω^* that is unknown to them. However, they know that a higher ω^* is more likely to yield a higher grade. In particular, assume the grade is given by*

[2]The assumptions are that u is twice continuously differentiable with: (i)$u_{ee} < 0$ and $u_e(\underline{e}\theta, \omega) > 0 > u_e(\bar{e}, \theta, \omega)$, (ii) $u_\theta, u_\omega > 0$ and (iii) $u_{e\theta} < 0$ and $u_{e\omega} > 0$. The direction of the derivatives is a normalization and the results would hold even when the signs are reversed.

$(\theta^* + a)\omega^*$.

The student must choose an effort level a, which determines their grade. For whatever the chosen effort level, the agent must pay a cost $c(a) = \frac{1}{2}a^2$. *And he repeats this process for infinitely many periods. Assume also that the student's prior is such that* $mathbbE[\omega] = \omega^*$ *and he is dogmatic about being of type* $\hat{\theta} > \theta^*$. [3] *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 $a_t^* = \omega^*$. *Nonetheless, because the agent does not know* ω^*, *he will choose* $a_t = \mathbb{E}_t(\omega)$ *where the expectation is taken with respect to the agent's belief at the beginning of period t. If he does not revise his effort choice for k periods, he will receive an average utility of* $(\theta^* + a_t^*)\omega^* - \frac{1}{2}a_t^{*2}$ *but he was expecting an average utility of* $(\hat{\theta}\theta + a_t^*)\omega^* - \frac{1}{2}a_t^{*2}$. *In response, he 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 agent to choose a lower effort at* $t + k$. *As a result, he will again receive an average utlitity that is lower than what he expected which will cause his belief to drift further down. This process will continue until the average utility equals his expected utility under the dogmatic belief that assigns probability 1 to* $\hat{\theta}$. *At that point, the stundent will have reached a self-defeating equilibrium and he will continue to choose sub-optimal effort forever.*

Although the model of a dogmatic modeler is able to rationalize 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 jump from one dogmatic belief to another. By doing so, the agent might end up

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

being dogmatic and correctly speciefied.

3.2 The Switcher

An agent is a *switcher* if they behave as a dogmatic, but is 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. Notice that their beliefs are still degenerate and assign probability one to a particular type, and zero to all other types. This means that a bayesian update on θ would cause no change in their beliefs. 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. They do so by keeping 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 off 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. 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.[4] 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.

[4]In a related problem @Schwarstein2021 proposes a similar updating procedure wich relies on the Bayesian hypothesis test.

By allowing the agent to keep track of the likelihoods and switching to an alternative type, the switcher can avoid the self-confirming equilibrium. However, if the prior belief on *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 this cases we say that the misspecified belief is persistent. Hence, in order to be able to determine which of these two models provides a better explanation of the observed behaviors, we must focus on cases in which the prior is diffused enough and the sequence of realized outcomes is such that the self-defeating equilibrium would not persist.

4 Framework 2

The agent is still of some type $\theta^* \in \Theta$ and the state is $\omega \sim F$. 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 ingreasing in θ^* and in ω. Whenever the agent observes a succes, he gets a payoff $v > 0$ and whenever the outcome is a failure, normalize the payoff to 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.

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

4.1 The Bayesian

A Bayesian agent simultaneously updates their beliefs about θ and ω by using Bayes' rule. The posterior odds at period t about θ after observing an outcome are given by:

$$\frac{p_t[\theta_H|\text{outcome}]}{p_t[\theta_M|\text{outcome}]} = \frac{p[\text{outcome}|\theta_H]p_{t-1}[\theta_H]}{p[\text{outcome}|\theta_M]p_{t-1}[\theta_M]}$$

and

$$\frac{p_t[\theta_M|\text{outcome}]}{p_t[\theta_L|\text{outcome}]} = \frac{p[\text{outcome}|\theta_M]p_{t-1}[\theta_M]}{p[\text{outcome}|\theta_L]p_{t-1}[\theta_L]}$$

Where p_{t-1} is the prior at period t and $p[\text{outcome}|\theta] = \sum_\omega p[\text{outcome}|\theta, \omega, e]p_{t-1}(\omega)$ is the probability of observing the outcome given the agent's type and the effort chosen. The update is symmetric for ω.

Bayesian agents always choose the effort level that maximizes their 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, although the Bayesian agent is the closest to a fully rational agent discussed here, they might not learn their true type. 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 with the results.

4.2 The Self-Serving Bayesian

A self-serving bayesian is an agent who uses a biased version of Bayes rule to update his beliefs.He will update his beliefs about the state ω and his type θ 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 greter degree than an unbiased agent would. To model the self-serving attribution bias, I take the approach of Benjamin [2019], where the posterior is given by:

$$p_t[\theta_H|\text{outcome}] = \frac{p[\text{outcome}|\theta]^{c_s^\theta \mathbb{I}\{\text{success}\}+c_f^\theta \mathbb{I}\{\text{failure}\}} p_{t-1}[\theta]}{\sum_{\theta' \in \Theta} p[\text{outcome}|\theta']^{c_s^\theta \mathbb{I}\{\text{success}\}+c_f^\theta \mathbb{I}\{\text{failure}\}} p_{t-1}[\theta']}$$

c_s^θ and c_f^θ are the self-serving attribution bias parameters for the agent's type θ. If $c_s^\theta = c_f^\theta = 1$, the agent is unbiased and the update is the same as the Bayesian update. On the other hand, if $c_s^\theta > c_f \theta$ the agent over-attributes success to their type and under-attributes failure to their type[5].

The update for ω is analogous but with c_f^ω and c_s^ω instead of c_f^θ and c_s^θ and the bias is present whenever $c_f^\omega > c_s^\omega$. That is, the agent over-attributes failure to a low state relative the higher states and under-attributes success to a low state relative to the higher states.

5 A Unifying Example

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 [6] 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 following table:

[5]notice that the values of c_s^θ and c_f^θ are not restricted to be greater than 1. If they are both equal to each other but less (more) than one, then the bias is simply underinference (overinference).

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

	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

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 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 direct comparison of all 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. [7] In what follows I explain the details of how this example was implemented in the lab.

6 Experimental Design

I recruited XXX undergraduate subjects from the CESS lab at NYU who participated in an in-person experiment. Sessions lasted approximately XXX hours and subjects earned an average payment of XXX. The experiment was programmed using oTree [**?**].

The experiment consisted of 2 treatments: the *ego-relevant* condition and the *stereotype* condition. Subjects participated in only one of the treatments. Treatments were randomly assigned at the session level. 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.

[7] because the setting does not match that of @Heidhues2018, there will be situationf for which the theory does not provie 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.

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 correspond 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 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 a signal for 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 part 3.

Randomness is controlled throughout the experiment and sessions by setting a seed at the beggining of the first session. The seed was drawn at random and remained fixed for all sessions. 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 and thus, if they use the same updating procedure, they should be choosing the same gambles. This design feature allows me to identify differences in updating procedures across subjects.

7 Analysis

Conclusion

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.

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.