WHAT YOU SEE IS ALL THERE IS*

BENJAMIN ENKE

News reports and communication are inherently constrained by space, time, and attention. As a result, news sources often condition the decision of whether to share a piece of information on the similarity between the signal and the prior belief of the audience, which generates a sample selection problem. This article experimentally studies how people form beliefs in these contexts, in particular the mechanisms behind errors in statistical reasoning. I document that a substantial fraction of experimental participants follows a simple "what you see is all there is" heuristic, according to which participants exclusively consider information that is right in front of them, and directly use the sample mean to estimate the population mean. A series of treatments aimed at identifying mechanisms suggests that for many participants, unobserved signals do not even come to mind. I provide causal evidence that the frequency of such incorrect mental models is a function of the computational complexity of the decision problem. These results point to the context dependence of what comes to mind and the resulting errors in belief updating. *JEL* Codes: D03; D80; D84.

I. INTRODUCTION

News reports and communication are both inherently constrained by space, time, and attention. As a result, news sources often condition the decision of whether to share a piece of information on the similarity between the signal and the prior belief of the audience. In some cases, news reports and communication disproportionately focus on events that are likely to move the audience's priors, such as the occurrence of terror attacks, large movements in stock prices, or surprising research findings. Although these types of events are routinely covered, the corresponding nonevents are not: one rarely reads headlines such as "No terror attack in Afghanistan today." In other cases, news providers

*I thank Andrei Shleifer and three extraordinarily constructive and helpful referees for comments that substantially improved the article. For discussions and comments, I am also grateful to Doug Bernheim, Thomas Dohmen, Armin Falk, Thomas Graeber, Shengwu Li, Josh Schwartzstein, Lukas Wenner, Florian Zimmermann, and seminar participants at Cornell, Harvard, Munich, Princeton, Wharton, BEAM 2017, SITE Experimental Economics 2017, and the 2016 ECBE conference. Financial support through the Bonn Graduate School of Economics and the Center for Economics and Neuroscience Bonn is gratefully acknowledged.

The Quarterly Journal of Economics (2020), 1363–1398. doi:10.1093/qje/qjaa012.
Advance Access publication on May 18, 2020.

supply news that align with people's priors but omit those that do not. For example, social networks like Facebook exclude stories from newsfeeds that go against users' previously articulated views. Regardless of the specific direction of the sample selection problem, these contexts share the feature that whether a specific signal gets transmitted depends on how this signal compares to the audience's prior. In the presence of such selection problems, people need to draw inferences from (colloquially speaking) "unobserved" signals.

While an active theoretical literature has linked selection problems in belief updating to various economic applications,[1] empirical work on people's reasoning in such contexts is more limited. Moreover, if people actually fail to take into account unobserved information, a perhaps even more fundamental question concerns the mechanisms behind such a bias. As reflected by a recent comprehensive survey paper on errors in statistical reasoning (Benjamin 2019), researchers have accumulated a broad set of reduced-form judgmental biases. Despite early calls for empirical work on the microfoundations of biases (Fudenberg 2006), relatively little is known about the mechanisms that underlie judgment errors. In the present context, a promising candidate mechanism is the idea that agents maintain an incorrect mental model of the environment because selection does not even come to mind when a decision is taken: people may never even ask themselves what it is that is not directly in front of them.

This articles tackles these two sets of issues—how people process selected information and the role of mental models therein—by developing a tightly structured individual decision-making experiment that operationalizes the selection problems discussed already. In the experiment, the entire information-generating process is computerized and known to participants. Subjects estimate an unknown state of the world and are paid for accuracy. The true state is generated as an average of six i.i.d. random draws from the simple discretized uniform distribution {50, 70, 90, 110, 130, 150}. I refer to these random draws as signals. Participants observe one of these six signals at random and subsequently indicate whether they believe the true state to be above or below 100. Thereafter, participants observe additional signals by interacting with a computerized information source. Just like in the

1. See Levy and Razin (2017), Han and Hirshleifer (2015), Jehiel (2018), and Jackson (2019).

motivating examples, this information source transparently conditions its behavior on the participant's first stated belief. On a participant's computer screen, the information source shares all signals that "align" with the participant's first stated belief (e.g., are smaller than 100 if the first belief is below 100) but not all signals that "contradict" the first belief (e.g., are larger than 100 if the first belief is below 100). Afterward, participants guess the state.

Bayesian inference would require participants to draw an inference about signals that do not appear on their computer screens, just like readers should infer something from the fact that a news outlet does not report on the stock market on a given day. In what follows, I colloquially say that participants "do not see" these latter signals, even though in an information-theoretic sense, this constitutes coarse information.

In a between-subjects design, I compare beliefs in this Selected treatment with those in a Control condition in which subjects receive the same objective information as those in Selected except that all signals physically appear on subjects' screens. Comparing beliefs across the two treatments allows me to causally identify participants' tendency to neglect selection problems in processing information, holding fixed the objective informational content of the signals.

The results document that beliefs exhibit large and statistically significant differences across the two treatments. Whenever participants' first signal is above 100, their final stated beliefs tend to be upward biased and conversely for initial signals below 100. I show that this pattern is robust against the provision of some feedback.

To disaggregate these cross-treatment differences, I analyze individual decision rules. Participants' responses are often heuristic in nature and reflect significant rounding to multiples of 5 or 10. Although individual decisions are noisy, these heuristics appear to have a systematic component. To identify this systematic part, the analysis estimates an individual-level parameter that pins down updating rules in relation to Bayesian rationality. Here, the distribution of updating types follows a bimodal structure: the modal responses of 60% of all participants are either Bayesian or reflect full neglect. In fact, even 87% of those participants that exhibit stable identifiable decision types can be characterized as exactly rational or exactly full neglect. Thus, a significant fraction of participants states beliefs whose stable component corresponds

to fully ignoring what they do not see and averaging the visible data.

Economists are increasingly interested in the mechanisms behind reduced-form errors in statistical reasoning, probably because of the view that this may help develop appropriate debiasing strategies or inform theoretical work. In the present context, the patterns are prima facie consistent with two alternative accounts of the data. A first is that—as posited in much recent theoretical work discussed below—neglect reflects an incorrect mental model of the data-generating process that arises because certain aspects of the problem do not even come to mind. Here, people may never even ask themselves which signals they do not see and why. Relatedly, a recent literature in cognitive psychology on the metaphor of the "naive intuitive statistician" argues that people are reasonably skilled statisticians but often naively assume that their information samples are unbiased and that sample properties can be directly used to estimate population analogs (Fiedler and Juslin 2006; Juslin, Winman, and Hansson 2007). According to this "incorrect mental models" perspective, the probability that selection comes to mind may be a function of the computational complexity of the decision problem. This is because decision makers need to allocate scarce cognitive resources between (i) setting up a mental model and (ii) computational implementation. Thus, a perspective of incorrect mental models suggests that people should be less likely to develop a correct mental model if they are cognitively busier with (or distracted by) computationally implementing a given solution strategy.

A plausible second view of the mechanisms behind neglect is that people are aware of the unobserved signals but struggle with the conceptual or computational difficulty of correcting for selection. To investigate the relative importance of these two accounts, I implement three sets of follow-up treatments. Each treatment variation predicts a change in behavior under only one of the two accounts.

First, I design a treatment in which the presence of a selection problem is eliminated, but subjects still need to process unobserved signals. If neglect was largely driven by the conceptual or computational difficulty of correcting for selection, then neglect should disappear in this treatment. Operationally, subjects observe four randomly selected signals, while four additional signals are not directly communicated to them. As in the baseline condition, participants do have information about the

unobserved signals, which in this case is their unconditional expectation. Nonetheless, a considerable fraction of subjects again follows a "what you see is all there is" heuristic of averaging the visible data. This shows that people struggle not (only) with conceptually thinking through a potential selection problem. Instead, they appear to have a more general tendency to estimate population means through sample means, where the "sample" is given by what is right in front of them and hence top of mind. As I discuss in Section VI, the averaging of observed data appears to tie together several recent findings in the experimental literature, including work in psychology on the "naive intuitive statistician."

As a second test between the two alternative mechanisms behind neglect, I devise treatments that hold the conceptual difficulty of accounting for selection constant but vary the cognitive resources that participants have at their disposal to set up a correct mental model. To this effect, I vary the computational complexity of computing beliefs in such a way that it plausibly affects only the probability that the unobserved signals come to mind. The experiments operationalize complexity in two different ways: the complexity of the signal space and the number of signals. First, to vary the complexity of the signal space, I implement treatments Complex and Simple. In Simple, the signal space is given by {70, 70, 70, 70, 70, 70, 130, 130, 130, 130, 130, 130}. In Complex, it is {70, 70, 70, 70, 70, 70, 104, 114, 128, 136, 148, 150}. In both treatments, whenever a participant states a first belief above 100, the selection problem can be overcome by remembering that an unobserved signal must be a 70. Thus, these treatments leave the conceptual and computational difficulty of accounting for selection constant (if the first belief is above 100). At the same time, these treatments vary the computational difficulty of computing a posterior belief and problem-induced cognitive load. Second, to manipulate the number of signals, participants in condition Few were confronted with the same signal space as those in Complex, but the true state was generated as the average of only two, (rather than six) random draws. Because all of these treatments fix the difficulty of backing out an unobserved signal, complexity can only matter to the extent that it induces cognitive load and reduces the probability that the unobserved signals come to mind.

The results show that increases in complexity (in terms of the number of signals and the complexity of the signal space) lead to substantially more neglect than in the respective comparison treatments. This is even though participants in the more complex

treatments work longer on the problems. The fact that variations in complexity matter for neglect even though the difficulty of accounting for selection is unchanged again highlights the role of (endogenous) incorrect mental models.

As a third test between the alternative mechanisms, I implement an experimental condition that includes a simple nudge on participants' decision screen to pay attention to, or remember, those signals that they "do not see." This intervention decreases neglect by about 50%, which again suggests that the unobserved signals otherwise did not come to subjects' minds in the first place.

In summary, the takeaways from the analysis of mechanisms are twofold. First, incorrect mental models play an important role in generating neglect. Unobserved signals do not seem to come to mind in the first place, which leads people to directly use the sample mean to estimate the population mean. Second, what comes to mind and the resulting mental models are not exogenously given "neglect parameters"—instead, they are context-dependent and endogenous to the computational complexity of the environment. These insights are potentially relevant not only for modeling updating errors but also for policy in terms of what will be an effective method to correct biased beliefs.

The article proceeds as follows. Section II describes the experimental design. Sections III–V present the results and study mechanisms. Section VI discusses related literature and offers concluding thoughts.

II. Experimental Design

II.A. Setup

The experiment was designed to achieve the following objectives: (i) full control over the data-generating process, (ii) exogenous manipulation of the degree of selection, (iii) a control condition that serves as a benchmark for updating without selected information, and (iv) incentive-compatible belief elicitation. Most importantly, a clean identification requires subjects' full knowledge of the data-generating process.

The main idea behind the design is to construct two sets of signals (two treatments) that result in the same Bayesian posterior, but only one information structure features a problem of selection. Subjects were asked to estimate an ex ante unknown state of the world θ and were paid for accuracy. The computer

TABLE I
OVERVIEW OF THE EXPERIMENTAL DESIGN

Stage 0	Stage 1	Stage 2	Stage 3	Stage 4
Computer determines state by drawing six signals	Subject receives one signal	First binary guess b_1 based on signal	Subject observes messages of information source	Continuous guess b_2

generated θ by drawing six times, with replacement, from the set $X = \{50, 70, 90, 110, 130, 150\}$. Draws from X are uniform. The average of these six draws then constituted the true state θ, which in the experiment is referred to as the "variable" that subjects needed to estimate. Henceforth, I refer to the random draws as signals.

In the course of the experiment, a subject interacted with a computerized information source that showed the subject (subsets of) the signals. An experimental task consisted of multiple stages, as summarized in Table I. First, after the computer generated the true state, a subject observed one randomly selected signal. Second, based on this first signal, subjects provided an incentivized guess b_1 about whether they believed θ to be smaller or larger than 100, $b_1 \in \{low, high\}$.[2]

Third, the information source showed the subject additional signals. This is the only stage in which treatments Selected and Control differed, as detailed below. Finally, after subjects observed the messages of the information source, they stated an incentivized belief about the state $b_2 \in [50, 150]$, with at most two decimals.

In Selected, the information source faced a budget constraint and hence conditioned its decision of which out of the remaining five signals to show the subject on the subject's first guess. Specifically, if the subject's first guess was higher than 100, the information source showed the subject all signals above 100, but at least three signals. Likewise, if the subject's first guess was smaller than 100, the information source showed the subject all signals below 100, but at least three signals. For example, if a participant's first guess was above 100 and only two of the remaining five signals were above 100, the information source showed

2. If the true state equaled 100, subjects received the full payment for either guess.

the subject these two signals and one randomly selected signal of those below 100. If four signals were above 100, the subject would be shown (only) these four. In what follows, I refer to the signals that the information source did not share with subjects as "unobserved" or as signals that subjects "do not see." This terminology is purely colloquial in nature and meant to make it salient that these signals do not appear on subjects' decision screens. In an information-theoretic sense, these "unobserved" signals constitute coarse information.

In summary, subjects in Selected faced a selection problem akin to the examples discussed in the introduction in that the information source conditions its messages (whether to send a signal) on the subject's prior. Given the simplified discretized uniform distribution over the signal space, it was rather straightforward for subjects to infer which types of signals were unobserved. Being sophisticated about selection requires subjects to understand that when they first guessed $b_1 = high$, an unobserved signal was 70, in expectation, while it was 130 when they first guessed $b_1 = low$.

Treatment Control was designed to deliver the same Bayesian posterior as Selected without the presence of a selection problem. In the Control condition, participants observed two types of signals on their decision screens. First, they observed those signals that subjects in the Selected treatment also observed. Second, they were also shown a coarse version of the signals that subjects in the Selected condition did not observe. Specifically, if an unobserved signal was in {50, 70, 90}, the information source communicated 70 to the subject, while if the unobserved signal was in {110, 130, 150}, the information source communicated 130.[3] These coarse messages equal the expected signal conditional on a subject's first guess in Selected. Thus, the informational content of the Selected and the Control treatments is identical.

Participants solved eight tasks with independent signal draws. To keep the experimental setup close to the motivating examples in which people need to process information about multiple variables of interest, the baseline experimental setup was such that subjects completed two tasks at the same time (on the same decision screen). In the instructions and in the computer program, this was referred to as estimating "variable A" and "variable B,"

3. On their computer screens, there was no way for subjects to tell apart a "realized" 70 and an "expected" 70. I made this design choice because telling them apart is redundant for rational inference.

respectively. Accordingly, subjects observed a first signal for each variable, then provided a first guess for each variable, and were then shown the subsequent messages of the information source, again for both variables. To avoid confusion, both the experimental instructions and the computer program specified which variable a signal belongs to by adding a capital letter. For example, subjects' first signals in the first period (the first two tasks) would be given by $A - 130$ and $B - 150$. This procedure was the same in Control and Selected. In total, subjects completed four periods (eight tasks), summarized in Table II. All subjects were exposed to the same sets of signal realizations. Below, I discuss a treatment that verifies that very similar results hold if subjects complete these eight tasks strictly sequentially.

The intrinsic interest of this study is in subjects' second guesses; the first guess only serves the purpose of imposing a selection problem akin to the examples described in the introduction. Thus, to reduce noise, the instructions mentioned that subjects' earnings from the first guess would be maximized in expectation if they followed the first signal, that is, stated a guess above (below) 100 if the signal was above (below) 100.

Control questions verified that subjects understood the process generating the data. For example, subjects were asked, "Assume that you issued a first guess of larger than 100. Which draws will the information source show you no matter what? (a) None. (b) Those above 100. (c) Those below 100." Only once subjects had correctly solved all control questions could they proceed to the experiment.[4] Online Appendix H contains the experimental instructions and control questions.

II.B. Theoretical Considerations

This subsection develops a simple, mechanical formal framework to fix ideas about the experimental design above. I use this framework below for model-based empirical analyses. The true state of the world is given by $\theta = \sum_{k=1}^{6} \frac{s_k}{6}$. Let $\mathbb{Z}(b_1)$ denote the set

4. The control questions followed a multiple-choice format with three to four questions per screen. Thus, trial-and-error was very cumbersome. Moreover, the BonnEconLab has a control room in which the experimenter can monitor the decision screens of all experimental subjects. Thus, whenever a subject appeared to have problems in answering the control questions, an experimenter approached the subject, clarified open questions (if any), and excluded the subject from the experiment if they did not appear to understand the instructions.

TABLE II

OVERVIEW OF THE EXPERIMENTAL TASKS

True state	First signal	Observed signal A	Observed signal B	Observed signal C	Observed signal D	Unobs. signal E	Unobs. signal F	Bayesian belief	Neglect belief
96.67	130	130	150	70	—	50	50	103.33	120.00
110.00	150	110	150	110	—	50	90	110.00	130.00
93.33	50	90	50	130	—	110	130	96.67	80.00
90.00	110	150	90	50	—	50	90	90.00	100.00
103.33	150	110	130	70	—	70	90	100.00	115.00
116.67	90	90	70	150	—	150	150	110.00	100.00
116.67	110	150	130	150	110	50	—	120.00	130.00
86.67	130	130	90	110	—	70	50	90.00	100.00

Notes: Overview of the belief formation tasks in order of appearance. The categorization into observed and unobserved signals applies to the case in which subjects follow their first signal, that is, guess $\geqslant 100$ if their signal was larger than 100, and $<$ otherwise. Subjects in the Selected treatment observed only their own signal and the "observed" signals. Subjects in the Control condition additionally had access to a coarse version of the "unobserved" signals, that is, if the corresponding signal was less than 100, they saw 70, and if the signal was larger than 100, they saw 130. See equations (1) and (2) for the Bayesian and neglect benchmarks.

of signals a subject actually sees on their computer screen, which depends on b_1. Denote $N = |\mathbb{Z}|$. Given a set of signals, a Bayesian would compute the mean posterior belief b_B as

$$b_B = \frac{\left[\sum_{k=1}^{N} s_{k\in\mathbb{Z}(b_1)}\right] + (6-N)\cdot \mathrm{E}[s_{k\notin\mathbb{Z}(b_1)} \mid b_1]}{6}, \tag{1}$$

where $s_k \in \mathbb{Z}(b_1)$ denotes a signal that appears on the decision screen. The second term in the numerator corresponds to the inference of a Bayesian of those signals that are not shown, which is the expectation conditional on the first belief.

I introduce theoretical benchmarks for neglect. A first possibility is that the agent applies a heuristic of "what you see is all there is" and does not draw any inferences from unobserved signals but just averages the observed data:

$$b_{N,1} = \frac{\sum_{k=1}^{N} s_{k\in\mathbb{Z}(b_1)}}{N}. \tag{2}$$

Comparing this benchmark with equation (1), we see that averaging the visible data generates two potential sources of error. First, the sample may be biased: because only $s_k \in \mathbb{Z}(b_1)$ appear in the numerator, b_1 determines whether predominantly high or low signals are taken into account. This is the traditional sample selection problem.

A second source of error arises because even if $\mathbb{Z}$ did not depend on b_1 (if there were no systematic sample selection), equation (2) would still ignore the unobserved signals. This is important because even if $\mathbb{Z}$ was determined at random, the decision maker has prior knowledge about the unobserved signals that he can make use of, which is that $E[s_k] = 100$.

A plausible alternative specification of a neglect benchmark eliminates the second type of error by positing that participants are aware of the signals they do not see but fail to understand the sample selection problem created in the process. Such a decision maker imputes the unconditional expectation of $E[s_k] = 100$ for any unobserved signal. The second neglect benchmark is given by

$$b_{N,2} = \frac{\left[\sum_{k=1}^{6} s_{k\in\mathbb{Z}(b_1)}\right] + (6-N)E[s_{k\notin\mathbb{Z}(b_1)}]}{6}. \tag{3}$$

It is perhaps helpful to provide an interpretation of the psychological difference between the two neglect benchmarks in equations (2) and (3). The agent in equation (3) only struggles with understanding (or computing) conditional expectations. The agent in equation (2) ignores the unobserved signals altogether, plausibly because he never actively thinks about how many signals there are. Because the unobserved signals are not top of mind, he naively uses the (visible) sample mean to estimate the population mean. Indeed, a long literature in cognitive psychology on the metaphor of a "naive intuitive statistician" posits that people have a tendency to directly use sample moments to estimate population analogs (Fiedler and Juslin 2006; Juslin, Winman, and Hansson 2007).

The main experiments were not designed to distinguish between these two neglect benchmarks. The correlation between $b_{N,1}$ and $b_{N,2}$ in my experimental tasks is $\rho = 0.99$, and they make quantitatively very similar predictions. However, in follow-up experiments (discussed in Section IV), I use the distinction between the two benchmarks to tease out the mechanisms behind neglect. The results show that a large majority of those subjects that are not Bayesian appear to follow the first neglect benchmark. Hence I use $b_{N,1}$ in what follows.[5]

Let $\chi \in [0, 1]$ parameterize the degree of neglect such that $\chi = 1$ implies full neglect. Then a decision maker's belief b can be expressed as a weighted average of b_B and $b_{N,1}$ plus decision noise ϵ:

$$\begin{aligned} b &= (1-\chi)b_B + \chi b_{N,1} + \epsilon \\ &= b_B + \chi \underbrace{\frac{6-N}{6}\left(\bar{s}_{k\in\mathbb{Z}(b_1)} - E[s_{k\notin\mathbb{Z}(b_1)}|b_1]\right)}_{\equiv d} + \epsilon \end{aligned} \tag{4}$$

$$= b_B + \chi d + \epsilon, \tag{5}$$

where $\bar{s}_{k\in\mathbb{Z}(b_1)}$ is the average visible signal and ϵ is a mean 0 random computational error. The systematic component of a subject's belief b can be expressed as Bayesian belief plus a distortion term

5. Table 3 in Online Appendix B and Figure 6 in Online Appendix C reproduce the main results using the $b_{N,2}$ benchmark. The results are almost identical to those presented below.

TABLE III
TREATMENT OVERVIEW

Treatment	# of subjects	Ave. earnings (euros)
Selected	74	12.77
Control	38	17.83
Sequential	75	11.28
Feedback	75	15.08
Random	75	12.10
Complex	75	14.28
Simple	75	14.47
Few	75	17.43
Nudge	72	12.18
Selected Replication	75	12.48

Notes: Horizontal lines indicate which treatments were randomized within the same experimental sessions. Payments included a show-up fee of €10 in Feedback and of €6 in all other treatments.

d times the neglect parameter χ. I use this formal framework to compute estimates of neglect $\hat{\chi}$ and decision noise $|\hat{\epsilon}|$.

II.C. Procedural Details

Apart from the treatments described above, I implemented eight additional treatments that will be discussed below. Table III provides an overview of all treatments; horizontal lines indicate which treatments were randomized within experimental sessions.

The experiments were conducted at the BonnEconLab of the University of Bonn and computerized using z-Tree (Fischbacher 2007). Participants were recruited using hroot (Bock, Baetge, and Nicklisch 2014). After the written instructions were distributed, subjects had 10 minutes to familiarize themselves with the task. Each period consisted of two computer screens. On the first screen, subjects were informed of the first signal and issued a binary guess. On the second screen, participants received the messages from the information source and stated a point belief. Sessions lasted 50 minutes on average.

All decisions were financially incentivized, in expectation: in total, subjects took 16 decisions, 1 of which was randomly selected for payment. This constitutes an incentive-compatible mechanism in such a setup (Azrieli, Chambers, and Healy 2018). The probability that a second (point) belief was randomly selected for payment was 90%, and one of the binary first guesses was chosen with probability 10%. The binary first guess was incentivized such that subjects received €18 if the guess was correct and

nothing otherwise. The continuous point beliefs were incentivized using a quadratic scoring rule with maximum variable earnings of €18, that is, variable earnings of subject i in task j equaled $\pi_i^j = \max\{0; 18 - 0.2 \times (b_i^j - \theta^j)^2\}$.

III. Results

III.A. Baseline Results

1. Preliminaries. The object of interest in the analysis is a potential treatment difference in the second beliefs that subjects state. For completeness, across the two treatments, 93% of all first binary guesses follow the first signal and enter a high (low) first guess if the first signal is above (below) 100. Online Appendix A presents a set of robustness checks that restrict the analysis to observations that followed the first signal.

2. Beliefs across Tasks. Table IV presents an overview of the results in each of the eight tasks. For ease of comparison, I provide the benchmarks of full neglect and Bayesian beliefs, respectively. Reassuringly, beliefs in the Control condition follow the Bayesian prediction very closely, suggesting that the experimental setup was not systematically misconstrued by subjects: in the absence of selected information, people state rational beliefs. In the Selected treatment, however, beliefs are distorted away from the Bayesian benchmark toward the full neglect belief. In all eight tasks, beliefs significantly differ between treatments at least at the 10% level, and usually at the 1% level (Wilcoxon ranksum tests).

3. Econometric Analysis. In the remainder of the article, treatment comparisons will be conducted by pooling the data across tasks for brevity and to eliminate potential multiple-testing concerns. Pooling the data requires transforming the beliefs data into a scale that has the same meaning across tasks. For this purpose, I make use of the simple belief formation rule introduced in Section II.B, which has the additional advantage that going forward, all estimated quantities will have direct theoretical counterparts. I use equation (5) to estimate the neglect implied in the belief of subject i in task j:

(6)
$$\hat{\chi}_i^j = E[\chi_i^j | b_i^j] = \frac{b_i^j - b_B^j}{d^j} = \frac{6(b_i^j - b_B^j)}{(6 - N^j)\left(\bar{s}^j_{k\in\mathbb{Z}(b_1)} - E[s^j_{k\notin\mathbb{Z}(b_1)} | b^j_{i,1}]\right)}.$$

TABLE IV

OVERVIEW OF BELIEFS ACROSS TASKS

True state (1)	First signal (2)	Bayesian belief (3)	Neglect belief (4)	Median belief Control (5)	Median belief Selected (6)	Mean belief Control (7)	Mean belief Selected (8)	p-value (ranksum) (9)
96.67	High	103.33	122.00	103.00	110.00	104.84	107.67	.0661
110.00	High	110.00	130.00	110.00	120.00	109.79	119.36	.0001
93.33	Low	96.67	80.00	96.50	90.00	96.58	90.88	.0130
90.00	High	90.00	100.00	90.00	90.00	90.21	94.00	.0536
103.33	High	100.00	115.00	100.00	110.00	98.79	107.78	.0001
116.67	Low	110.00	100.00	110.00	110.00	110.29	108.08	.0635
116.67	High	120.00	130.00	120.00	123.00	118.29	122.04	.0099
86.67	High	90.00	100.00	90.00	90.00	89.16	95.89	.0022

Notes: Overview of the estimation tasks in order of appearance. See Table II for details on the signals in each task as well as the computation of the Bayesian and full neglect benchmarks. High (low) private signals are defined as signals above (below) 100. The p-value refers to a Wilcoxon ranksum test between beliefs in Selected and Control.

Note that this analytical tool corresponds to a simple linear transformation of the raw beliefs data (subtract the Bayesian belief and divide by the distortion term d, which is only a function of the signal realizations). This method only converts the data into a consistent interval, so that subjects' beliefs (i) are on the same scale across tasks and (ii) can be directly interpreted as reflecting Bayesian ($\hat{\chi} = 0$), full neglect ($\hat{\chi} = 1$), or intermediate levels.

Although $\hat{\chi}_i^j$ should in principle be between 0 and 1, in the experimental data naturally not all observations lie within this interval, likely at least partly due to typing mistakes and random computational errors. This produces outliers that are partly severe. Across the treatments in Table III ($N = 5{,}416$ belief statements), the minimum implied $\hat{\chi}_i^j$ is -21 and the maximum 12.7. To avoid arbitrary exclusion criteria while dealing with outliers, throughout the article I present three different sets of regression specifications. First, I present an analysis with median regressions that includes the full sample of beliefs, including large outliers. Second, I perform an OLS analysis in which I winsorize the data at $|\hat{\chi}_i^j| = 3$. That is, I replace each belief that is larger (smaller) than 3 (-3) by the corresponding value. This affects 3% of all observations. Third, I present an OLS analysis on a trimmed sample, where I drop all observations with $|\hat{\chi}_i^j| > 3$. For completeness, Online Appendix A presents an additional set of specifications in which I implement OLS regressions on the full sample, including all outliers. The results are similar to those reported in the main text.

Table V presents the results. In these analyses, the unit of observation is a subject-task, for a total of usually eight observations per subject.[6] The standard errors are clustered at the subject level. All regressions include experimental session fixed effects, leveraging random assignment into treatments within sessions.

The results confirm a large and statistically significant aggregate treatment difference between Control and Selected. In column (1), the median regression only controls for session fixed effects. Column (2) adds a vector of controls: fixed effects for each experimental task interacted with the first guess (high / low) of the subject, as well as controls for individual characteristics. In columns (3) and (4), the dependent variable is winsorized at |3|, and I estimate OLS regressions. In columns (5) and (6), the sample

6. In a few cases, subjects did not enter a belief on time, so these observations are missing.

TABLE V

BASELINE RESULTS: TREATMENTS SELECTED AND CONTROL

	Dependent variable: Neglect $\hat{\chi}_i^j$					
	Median regression		OLS winsorized		OLS trimmed	
	(1)	(2)	(3)	(4)	(5)	(6)
0 if Control, 1 if Selected	0.40***	0.50***	0.54***	0.60***	0.51***	0.54***
	(0.08)	(0.10)	(0.09)	(0.09)	(0.09)	(0.09)
Session FE	Yes	Yes	Yes	Yes	Yes	Yes
Task FE × prior	No	Yes	No	Yes	No	Yes
Controls	No	Yes	No	Yes	No	Yes
Observations	894	894	894	894	874	874
R^2	0.07	0.10	0.09	0.11	0.10	0.11

Notes: Regression estimates, with robust standard errors (clustered at subject level) in parentheses. The dependent variable is the neglect $\hat{\chi}_i^j$ that is implied in a given belief. The sample includes each of subjects' eight beliefs in the Selected and Control conditions. Columns (1) and (2) report median regressions, and columns (3)–(6) are OLS regressions. In columns (3) and (4), the dependent variable is winsorized at $|\hat{\chi}_i^j| = 3$. In columns (5) and (6), the sample is trimmed at $|\hat{\chi}_i^j| = 3$. Controls include gender, high school grades, and log monthly disposable income. * $p < .10$, ** $p < .05$, *** $p < .01$.

excludes observations with $|\hat{\chi}_i^j| > 3$. Throughout, the coefficient is quantitatively large and suggests that—relative to the control treatment—subjects in Selected exhibit a neglect of 0.4–0.6 units of χ.

The bias implies lower earnings of subjects in the Selected condition. The expected profit from all eight belief formation tasks is €6.33 in Selected and €10.32 in Control. Actual profits, which include a show-up fee and depend on a random draw, are €17.56 ($20) in Control and €12.73 ($15) in Selected.

III.B. Robustness Treatments

1. Sequential Tasks. To assess the extent to which the simultaneous presentation of two variables induces neglect, I implemented treatment Sequential. This treatment was randomized along with Control and Selected within experimental sessions. Sequential is identical to Selected, except that all eight tasks were presented in eight, rather than four, consecutive rounds. Online Appendix D discusses the results from this treatment in detail. Overall, the results are very similar to those in Selected. To illustrate, Figure I plots the median and mean $\hat{\chi}_i^j$ across treatments, along with standard error bars. Although the median neglect

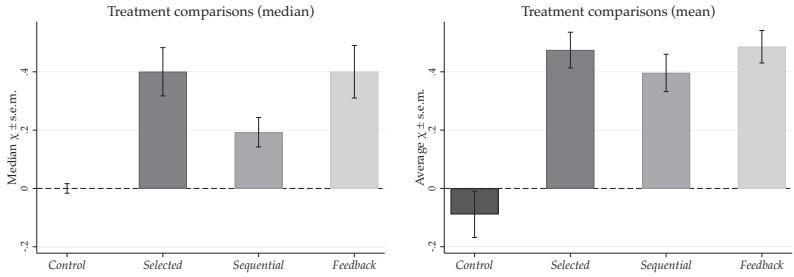

FIGURE I

Overview of Neglect $\hat{\chi}_i^j$ across Treatments

The left panel shows the median $\hat{\chi}_i^j$ across all subject-task observations. The right panel shows the average $\hat{\chi}_i^j$ across all subject-task observations, where as in Table V, columns (3) and (4), the data are winsorized at $|\hat{\chi}_i^j| = 3$. For treatment Feedback, the sample median and average are computed for the last eight beliefs to keep the results comparable to the other treatments. Standard error bars are computed based on clustering at the subject level.

estimate is significantly lower in Sequential than in Selected, the averages are very similar ($\bar{\hat{\chi}}_i^j = 0.49$ in Selected and $\bar{\hat{\chi}}_i^j = 0.42$ in Sequential). Moreover, neglect in Sequential is significantly higher than in Control.[7]

2. Feedback. A relevant question is whether people learn about their errors through feedback. In treatment Feedback, subjects first solved six tasks (again, two per period) that had the same structure as those in Selected but different signal realizations. Then they completed the same eight tasks as subjects in Selected. Thus, I can compare beliefs across treatments for the same tasks, yet subjects in Feedback have already completed six tasks and received feedback on them. After each period, subjects received feedback about their performance: (i) they were reminded of their continuous belief statement; (ii) they were informed of the corresponding true state; and (iii) they received information on the profits that would result from the respective task in case it would

7. As documented in Table 8 in Online Appendix D, median and average neglect are consistently lower in Sequential than in Selected. Although these differences are usually not statistically significant, they provide some very tentative evidence that the simultaneous presentation of problems induces higher cognitive load, which in turn increases neglect. See Section IV for a dicussion along these lines.

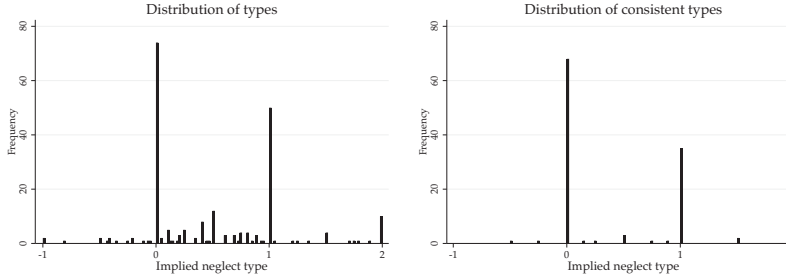

FIGURE II

Distribution of Modal Neglect Types $\hat{\chi}_i$ in Treatments Selected, Sequential, and Feedback

The left panel shows the distribution of all estimated neglect types, and the right panel the distribution of neglect types for which at least three beliefs are type-consistent (53% of all subjects). For belief j to be consistent with the estimated type means that $|\hat{\chi}_i - \hat{\chi}_i^j| \leqslant 0.05$.

be selected for payment. Online Appendix E provides a detailed analysis of the data. The data show no indication that feedback reduces the amount of neglect. Figure I illustrates this result.

III.C. Decision Rules and Heterogeneity Analysis

1. Type Distribution. To examine the subject-level distribution of neglect, I seek to identify a subject's neglect type $\hat{\chi}_i$, that is, an estimate of a subject's solution strategy, net of computational errors and heuristic rounding. For this purpose, for each subject i and candidate type $t \in \{-1, -0.9, \ldots, 2\}$, I count how many of the implied $\hat{\chi}_i^j$ (see equation (6)) satisfy $|t - \hat{\chi}_i^j| \leqslant 0.05$. Then, I classify each subject as $\hat{\chi}_i = t_{\max}$, where $t_{\max}$ is the candidate type that rationalizes the largest number of beliefs (see Fragiadakis, Knoepfle, and Niederle 2016 for a similar approach).[8]

The left panel of Figure II presents a histogram of these modal neglect types $\hat{\chi}_i$ in treatments Selected, Sequential, and Feedback. The data reveal a bimodal type distribution: 60% of all subjects are best characterized as Bayesian ($\hat{\chi}_i = 0$) or full neglect ($\hat{\chi}_i = 1$). For example, of those 150 subjects that are not approximately rational, one third (51) are classified as exactly or almost exactly

8. If more than one type rationalizes the maximal number of beliefs, I compute the average across t.

full neglect types ($0.95 \leqslant \chi_i \leqslant 1.05$).[9] In contrast, in treatment Control, 80% of all subjects are classified as exactly $\hat{\chi}_i = 0$; see Figure 7 in Appendix C.

2. Across-Task Consistency and Heuristic Responses. Because the left panel of Figure II shows modal types, the figure does not take into account the within-subject-across-task consistency in stated beliefs. To address this, I look at the number of beliefs that are consistent with a subject's modal type, where type-consistent means that the neglect parameter implied by a belief statement is close to the overall estimated type: $|\hat{\chi}_i - \hat{\chi}_i^j| \leqslant 0.05$. Figure 1 in Online Appendix C shows a histogram of the number of type-consistent beliefs. The average and median number of type-consistent beliefs are 3.2 and 3. The noisiness of the data in bounded rationality experiments—and the fact that a considerable fraction of subjects does not appear to behave according to a stable type—has recently been highlighted by Fragiadakis, Knoepfle, and Niederle (2016).[10] They and Costa–Gomes and Crawford (2006) propose that a subject should be viewed as having a stable type if at least 40% of their experimental actions are type-consistent.

The right panel of Figure II shows the distribution of modal neglect types, restricting attention to those 53% of all subjects for which at least three beliefs (∼40%) are type-consistent. The two spikes at $\chi_i = 0$ and $\chi_i = 1$ largely remain, yet the vast majority of all types $\hat{\chi}_i \neq 0, 1$ are relatively inconsistent across tasks. A perhaps remarkable 87% of those subjects who exhibit stable identifiable decision types can be characterized as exactly rational or exactly full neglect. Very few subjects exhibit a stable decision type of partial adjustment from neglect.

9. Figure 5 in Online Appendix C plots a histogram of the subject-task-specific $\hat{\chi}_i^j$, that is, the underlying raw beliefs data. Naturally, this distribution is noisier but also bimodal with spikes at 0 and 1.

10. The noisiness of the beliefs data appears to be at least partly driven by heuristic rounding to the nearest multiple of 5 or 10, akin to the patterns documented in a large survey literature on subjective expectations about economic variables (Manski 2004). In my data, 69% of reported beliefs are multiples of 10 and 84% are multiples of 5. These numbers are probably inflated because the Bayesian or full neglect benchmarks are themselves usually multiples of 5 or 10, compare Table II. Yet when I exclude beliefs that correspond to the Bayesian or full neglect benchmarks, still 53% are multiples of 10 and 75% multiples of 5.

In the full sample of subjects, for the $\hat{\chi}_i = 0$ types, 4.5 beliefs are type-consistent, on average. For the $\hat{\chi}_i = 1$ types, 3.4 beliefs are type-consistent, on average. However, for all types $\hat{\chi}_i \neq 0, 1$, the average number of type-consistent beliefs is only 2.0. Overall, these patterns suggest that across-task consistency is relatively low, in particular for the $\hat{\chi}_i \neq 0, 1$ types.[11] Still, to the extent that there is within-subject consistency in my data, it points to the presence of two fundamentally different updating types.

As a final remark on within-subject consistency, it is worth pointing out that the relatively inconsistent subjects are not just random noise around the rational benchmark. The average and median task-level implied neglect parameters of relatively inconsistent subjects are $\chi_i^j = 0.35$ and $\chi_i^j = 0.40$. This shows that the inconsistent types do neglect selection—just in a quantitatively inconsistent fashion across tasks.

3. Correlates of Neglect. Table 4 in Online Appendix B investigates the correlates of neglect in treatments Selected, Sequential, and Feedback. I find that better high school grades and longer response times are both negatively correlated with neglect. The quantitative magnitude of the relationship between response times and neglect is small. Interpreted causally, the regression coefficients suggest that response times would have to increase by about four minutes per task to move a full neglect belief to a Bayesian belief. However, the average response time in the data in the three treatments that are considered here is only 48 seconds, and it is 52 seconds in treatment Control. These magnitudes suggest that the type heterogeneity is not merely the result of the neglect types being lazier than the rational types.

IV. Mechanisms

IV.A. Framework

Understanding the mechanisms behind errors in statistical reasoning is likely to be relevant not only for theorists who are interested in formalizing and endogenizing people's errors but also for policy in terms of what will be an effective method to

11. The intracorrelations between modal, median, and average neglect types are all between 0.75 and 0.92. Figures 2–4 in Online Appendix C present histograms of (i) median subject-level neglect, (ii) average neglect parameters, and (iii) the subject-level standard deviation of implied neglect parameters.

correct biased beliefs. To structure the analysis, I pit two hypotheses against each other.

1. Theory A: Incorrect Mental Model. Participants have an initial mental default model according to which the unobserved signals are not top of mind. This default model could result from intuitive system 1 reasoning (Kahneman 2011), or it could be retrieved from memory as the "normal" version of a class of problems that people know how to solve (Kahneman and Miller 1986). If the unobserved signals do not come to mind, participants directly use the (visible) sample mean to estimate the population mean, akin to the psychological metaphor of a naive intuitive statistician who directly uses sample moments to estimate population analogs (Fiedler and Juslin 2006; Juslin, Winman, and Hansson 2007). This simple averaging process may be loosely summarized as "what you see is all there is."

If selection does come to mind, the participant reasons about whether and how it needs to be corrected for. Whether this happens partly depends on how the decision maker allocates cognitive resources between (i) setting up a mental model and (ii) computational implementation. In particular, people should be less likely to develop a correct mental model if they are cognitively busier with (or distracted by) computationally implementing a given solution strategy.

Linking this account to the literature, the importance of incorrect mental models is highlighted by an active theoretical literature (e.g., Jehiel 2005; Eyster and Rabin 2010; Schwartzstein 2014; Gabaix 2014; Esponda and Pouzo 2016; Spiegler 2016; Bohren and Hauser 2017; Heidhues, Kőszegi, and Strack 2018; Gagnon-Bartsch, Rabin, and Schwartzstein 2018). For example, the model in Spiegler (2017) focuses on how an agent naïvely extrapolates from partial data, which is reminiscent of the sample selection problem in this paper. Indeed, incorrect mental models are often implicitly, and sometimes explicitly, motivated and modeled as resulting from attentional processes (Gennaioli and Shleifer 2010).

2. Theory B: Conceptual or Computational Difficulty of Accounting for Selection. Participants are aware of the signals they do not see but struggle with the conceptual or computational difficulty of correcting for selection.

It is worth highlighting that these two stories are not necessarily mutually exclusive but relate to two distinct steps of a sequential reasoning process. In the first step, it is determined whether selection (the unobserved signals) are top of mind. In the second step, the decision maker reasons about how to correct for selection, if it comes to mind in the first place. In principle, it is conceivable that selection does not come to mind, but even if it did come to mind, the participant wouldn't be able to account for it.

The experiments that follow test the relative importance of these two stories by exogenously manipulating parameters that should lead to changes in reported beliefs according to one theory but not the other. I conduct three such comparative statics exercises:

(i) Holding fixed the presence of unobserved signals, I eliminate the presence of the selection problem. If neglect is largely driven by theory B, it should disappear in this treatment. If neglect is largely driven by theory A, it should remain roughly constant.
(ii) Holding fixed the conceptual and computational difficulty of accounting for selection, I increase the computational complexity of following a "what you see is all there is" averaging heuristic. If neglect is largely driven by theory B, then such complexity variations should have no effect. Under theory A, higher computational complexity should increase neglect because the decision maker is "distracted" by computational implementation and thus devotes less resources to thinking about what is not top of mind or visible.
(iii) Holding fixed the conceptual and computational difficulty of accounting for selection, I exogenously draw participants' attention to the unobserved signals. If neglect is largely driven by theory A, it should substantially decrease. If neglect is largely driven by theory B, it should remain constant.

IV.B. Eliminating the Selection Problem

1. Experimental Design. To test comparative statics prediction i, I implemented treatment Random. Random closely follows treatment Selected. The true state to be estimated now consists of the average of eight random draws from the same simple

discretized uniform distribution as before.[12] Deviating from the procedure in Selected, in Stage 3 of the experiment, a subject observed three signals that were selected at random, rather than based on a subject's first guess. The timeline of this treatment was otherwise identical to that in treatment Selected. In this setup, the Bayesian belief is given by

$$b_B = \frac{\left[\sum_{k=1}^{4} s_{k\in\mathbb{Z}(b_1)}\right] + 4 \cdot E[s_{k\notin\mathbb{Z}(b_1)}]}{8}, \tag{7}$$

while a "what you see is all there is" benchmark is given by the same equation as before:

$$b_{N,1} = \frac{\sum_{k=1}^{4} s_k}{4}. \tag{8}$$

It is worth pointing out that this treatment also directly speaks to the two potential neglect benchmarks for treatment Selected discussed in Section II.B: (i) a decision rule that assumes that subjects completely ignore information that is not visible on their computer screen and (ii) a decision rule that posits that participants are aware of the signals they do not see but wrongly assign them their unconditional rather than conditional expectation. If (ii) was the empirically correct benchmark, then subjects in Random should state Bayesian beliefs.

2. Results. The results are described in detail in Online Appendix F. To summarize, behavior in this treatment is very similar to behavior in treatment Selected. I again compute implied subject-level neglect parameters χ_i, where 0 corresponds to the Bayesian and 1 to the full neglect benchmark noted above. As shown in Figure III, the distribution of stated beliefs is again bimodal, with subjects either fully neglecting what they don't see or behaving rationally. Indeed, as shown in Online Appendix F, the distribution of neglect in this treatment is statistically indistinguishable from the one in treatment Selected.

12. In this treatment, the true state was determined as the average of eight (rather than six) random draws to allow for a larger number of invisible signals. With only two invisible signals, the Bayesian and full neglect benchmarks would have been too close to each other to allow for robust analyses that distinguish between these two updating types.

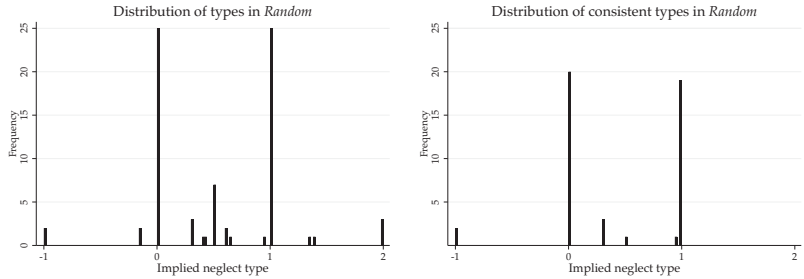

FIGURE III

Distribution of Modal Neglect Types $\hat{\chi}_i$ in Treatment Random

See the main text for the specification of the Bayesian and full neglect benchmarks. The left panel shows the distribution of all estimated neglect types, and the right panel the distribution of neglect types for which at least three beliefs are type-consistent (61% of all subjects). For belief j to be consistent with the estimated type means that $|\hat{\chi}_i - \hat{\chi}_i^j| \leqslant 0.05$.

I view the results of this treatment as suggesting two implications. First, a "what you see is all there is" heuristic describes behavior better than a theoretical benchmark in which subjects actively impute unconditional expectations for unobserved signals. This suggests that at least a majority, and probably a large majority, of those subjects that are classified as "neglect" types in treatment Selected do not at all take into account the unobserved signals. Second, the psychological mechanism behind neglect is probably not (just) a conceptual misunderstanding of selection problems but instead a general incorrect mental model according to which the unobserved signals do not come to mind in the first place.

IV.C. Computational Complexity as Distraction

1. Experimental Design. Next, I study how computational complexity affects selection neglect, in particular how it might induce cognitive load and hence distract participants from the unobserved signals. The experiments below exogenously manipulate the computational complexity of the updating problem but hold fixed the difficulty of accounting for selection itself. This thought experiment has the attractive feature that it narrows down the pathways through which complexity can affect belief updating: if the difficulty of correcting for selection remains unchanged, then differences in belief updating can plausibly be attributed to an

effect of computational complexity on how participants approach the problem (develop a mental model) in the first place. Given the absence of a general theory of what is complex, the experiments operationalize computational complexity in two different and arguably intuitive ways: (i) the complexity of the signal space and (ii) the number of signals in a given updating problem.

Complexity I: The Complexity of the Signal Space. To exogenously vary the complexity of the signal space, I conducted two treatments, Complex and Simple. These two treatments were both identical to treatment Selected except that the set of numbers from which the true state was determined was varied. In Complex, the signal space was given by

$$\{70, 70, 70, 70, 70, 70, 104, 114, 128, 136, 148, 150\}.$$

In Simple, it was

$$\{70, 70, 70, 70, 70, 70, 130, 130, 130, 130, 130, 130\}.$$

These two treatments are identical in a number of ways: (i) the prior is 100; (ii) the conditional expectations of being above and below 100 are 130 and 70, respectively; (iii) most important, they leave the difficulty of accounting for selection constant if subjects state a first guess of above 100 (i.e., in practice, when they receive a first signal above 100). In such cases, accounting for selection only requires subjects to notice (remember) that they are missing a few 70s on their decision screens. Thus, in both treatments, people's potential problems in computing conditional expectations cannot drive any results. For example, in one task, subjects in Complex observed 150, 104, 148, 114 on their decision screens, whereas those in Simple observed 130, 130, 130, 130.

Complexity II: The Number of Signals. Treatment Few was identical to Complex in almost all dimensions. The only difference is the number of random draws (signals) that determined the true state and were shown to subjects. In Few, the state was determined as the average of two, rather than six, random draws.

Subjects in Few also observed a first signal and then issued a first binary guess. Given that there are only two signals in total in this treatment, subjects then potentially observed one more signal from the information source. Subjects only observed this second

signal if it was above 100 and the subject's first guess was above 100, or if the second signal was below 100 and the subject's first guess below 100. Thus, in many tasks, subjects did not receive an additional (second) signal from the information source on the second decision screen. Notice that if subjects observe both signals, there is no selection problem, so that by design, the analysis of Few has to exclude the three experimental tasks for which this was the case.

Comparing treatments Few and Complex leaves the signal space and hence the difficulty of backing out unobserved signals unchanged. Still, the computational complexity of computing posteriors differs across treatments. For example, in one task, subjects in Complex observed 150, 104, 148, 114 on their decision screens, while those in Few observed 150.

In summary, all treatments hold the diffculty of accounting for selection constant but vary the computational burden of computing beliefs. A notable difference to earlier cognitive load experiments is that here cognitive load arises endogenously as a feature of the decision problem, rather than being exogenously induced by the experimenter.

Finally, note that comparing treatments Simple and Few is not meaningful by design because these treatments differ in two dimensions in ways that operate in opposite directions. Treatment Simple is simpler than Few in that is has a simpler signal space, but treatment Few is simpler in that it features a smaller number of signals. Thus, the analysis compares Complex to Simple and Complex to Few. Treatments Complex, Simple, and Few were all randomized within the same experimental sessions; compare Table III. Tables 5 and 6 in Online Appendix B show the signal realizations in these treatments.

2. Manipulation Checks. Given that the treatment variations here are arguably relatively subtle and do not have immediate antecedents in the literature, it is worth performing a manipulation check to verify that the computational complexity is indeed meaningfully higher in Complex than in Simple and Few. To provide such evidence, I consider data on (i) response times and (ii) the noisiness of responses across tasks. Higher computational complexity should translate into (i) longer response times and (ii) beliefs data that are noisier, or less consistent across tasks. Following equation (6), I estimate decision noise by comparing a subject's belief in task j with the belief they "should have"

stated given their estimated overall type $\hat{\chi}_i$: $|\hat{\epsilon}_i^j| = |\hat{\chi}_i^j - \hat{\chi}_i|$, where $\hat{\chi}_i$ is the overall estimate of i's type across tasks as derived in Section III.

Table 7 in Online Appendix B shows that both response times and decision noise are indeed significantly lower in Simple and Few, as compared with Complex. This provides reassuring evidence that the treatment variations actually induced meaningful variations in computational complexity as perceived by the experimental participants.[13]

3. Results. By design of the experiment, the analysis is restricted to those tasks in which subjects' first signal was above 100 so that any unobserved signal had to be a 70 in all treatments. Figure IV plots median and average levels of $\hat{\chi}_i^j$ across treatments. Here, just like in the regression tables, $|\hat{\chi}_i^j|$ is winsorized at 3 when I compute treatment averages. As predicted, treatment Complex generates substantially higher levels of neglect than do Simple and Few. The median implied neglect in Simple and Few is 0, though the averages are strictly positive ($\bar{\hat{\chi}}_i^j = 0.13$ in Simple and $\bar{\hat{\chi}}_i^j = 0.22$ in Few).

Table VI provides a set of corresponding regression analyses. In all regressions, the omitted baseline category is treatment Complex. By including treatment dummies for Simple and Few, the regressions compare Complex with Simple and Complex with Few.

Both treatment dummies have negative coefficients that are statistically significant. These results hold both in the analysis with median regressions (columns (1) and (2)) and in robustness checks in which the dependent variable is winsorized or trimmed (columns (3)–(6)). In terms of quantitative magnitude, the coefficients suggest that both types of complexity reductions caused a reduction in neglect by about 0.2−0.3 units of $\hat{\chi}_i^j$. Thus, the

13. A potential issue with the interpretation that higher computational complexity increases decision noise is that it is impossible for me to formally disentangle the story that decision error is lower for less computationally complex tasks from a scenario where decision error conditional on type is independent of computational complexity, but the more complex treatment changes the type distribution and decision errors are larger for neglect types. However, this alternative interpretation of the results is less plausible because the calculations that are required to be Bayesian are unambiguously more complicated than those required to follow the neglect benchmarks.

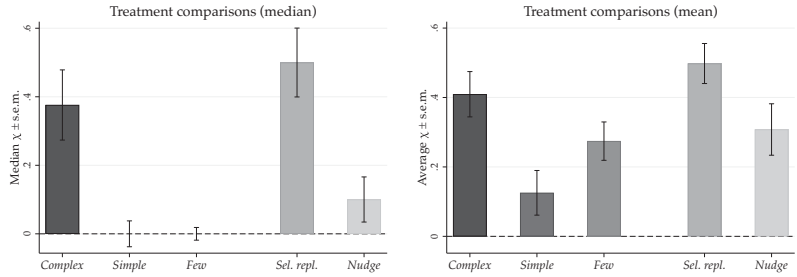

FIGURE IV

Overview of Neglect $\hat{\chi}_i^j$ across Treatments

The left panel shows the median $\hat{\chi}_i^j$ across all subject-task observations. The right panel shows the average $\hat{\chi}_i^j$ across all subject-task observations, whereas in columns (3) and (4) of Table V the data are winsorized at $|\hat{\chi}_i^j| = 3$. Standard error bars are computed based on clustering at the subject level. As explained in the text, by design, the analysis of treatments Complex, Simple, and Few is restricted to those experimental tasks in which the first signal was above 100. Moreover, also by design, for treatment Few the analysis excludes those tasks in which subjects observed both (and hence all) signals, so no selection problem was present. As explained in the main text, these data exclusions follow mechanically from the construction of the different treatments.

increased cognitive load from the computational stage of the problem appears to have systematic effects on how participants approach the conceptual stage of forming a mental model to begin with. This provides further evidence that in this context selection neglect is not (just) driven by the conceptual or computational difficulty of accounting for selection—as this was held constant across treatments—but by an incorrect mental model.

IV.D. Nudge Evidence

1. Experimental Design. If it is true that participants in Selected entertain an incorrect mental model, then nudging their attention toward (or reminding them of) the existence of the selection problem might attenuate the bias. Specifically, treatment Nudge was identical to Selected, except that both the end of subjects' written instructions and their decision screens contained the following hint:

"HINT: Also pay attention to those randomly drawn balls that are not shown to you by the information source."

TABLE VI
TREATMENTS COMPLEX, SIMPLE, AND FEW

	Dependent variable: Neglect $\hat{\chi}_i^j$					
Omitted category: Complex	Median regression		OLS winsorized		OLS trimmed	
	(1)	(2)	(3)	(4)	(5)	(6)
1 if Simple	−0.29**	−0.26***	−0.28***	−0.27***	−0.25***	−0.25***
	(0.12)	(0.10)	(0.09)	(0.09)	(0.08)	(0.09)
1 if Few	−0.29**	−0.24**	−0.17*	−0.29***	−0.18**	−0.22***
	(0.12)	(0.10)	(0.09)	(0.09)	(0.08)	(0.08)
Session FE	Yes	Yes	Yes	Yes	Yes	Yes
Task FE × prior	No	Yes	No	Yes	No	Yes
Controls	No	Yes	No	Yes	No	Yes
Observations	1,177	1,177	1,177	1,177	1,138	1,138
R^2	0.01	0.02	0.03	0.08	0.04	0.06

Notes: Regression estimates with robust standard errors (clustered at subject level) in parentheses. The sample includes treatments Complex, Simple, and Few. By the design of the experiment, the sample is restricted to those tasks in which following the first signal implies a first guess above 100. In treatment Few, experimental tasks in which subjects observe both signals are necessarily excluded because there is no scope for neglecting selection. Columns (1) and (2) report median regressions, and all other columns OLS regressions. In columns (3) and (4), $|\hat{\chi}_i^j|$ is winsorized at 3. In columns (5) and (6), the sample is trimmed at $|\hat{\chi}_i^j| = 3$. Controls include gender, high school grades, and log monthly disposable income. * $p < .10$, ** $p < .05$, *** $p < .01$.

Treatment Nudge was implemented along with a replication of treatment Selected to facilitate within-session randomization of subjects into treatments.[14]

2. Results. Figure IV shows that treatment Nudge generates lower levels of neglect than Selected Replication. Table VII provides a set of corresponding regression analyses. Treatment Nudge reduces neglect by about 0.2–0.4 units of $\hat{\chi}_i^j$, which corresponds to about half of the treatment difference between

14. To investigate whether subjects are capable of computing the conditional expectations that are required in the present experiment, treatments Selected and Sequential contained two incentivized follow-up questions: "Suppose you knew that ten balls were randomly drawn and that all of these balls had numbers GREATER than 100. What would you estimate is the average of these ten numbers?" Subjects were asked the same question with GREATER replaced by SMALLER. For each question, subjects received €0.50 for a correct response and €0.20 if the response was within 5 of the correct response. Figure 8 in Online Appendix C presents histograms of subjects' responses to these two questions. A large majority (almost 80%) of subjects guess the correct conditional expectations.

TABLE VII
TREATMENTS SELECTED REPLICATION AND NUDGE

	Dependent variable: Neglect $\hat{\chi}_i^j$					
	Median regression		OLS winsorized		OLS trimmed	
	(1)	(2)	(3)	(4)	(5)	(6)
0 if Selected Repl., 1 if Nudge	−0.40***	−0.20**	−0.20**	−0.21**	−0.22***	−0.24***
	(0.11)	(0.08)	(0.09)	(0.09)	(0.08)	(0.08)
Session FE	Yes	Yes	Yes	Yes	Yes	Yes
Task FE × prior	No	Yes	No	Yes	No	Yes
Controls	No	Yes	No	Yes	No	Yes
Observations	1,174	1,174	1,174	1,174	1,154	1,154
R^2	0.02	0.10	0.03	0.11	0.03	0.10

Notes: Regression estimates with robust standard errors (clustered at subject level) in parentheses. The sample includes treatments Selected Replication and Nudge. Columns (1) and (2) report median regressions, and all other columns OLS regressions. In columns (3) and (4), $|\hat{\chi}_i^j|$ is winsorized at 3. In columns (5) and (6), the sample is trimmed at $|\hat{\chi}_i^j| = 3$. Controls include gender, high school grades, and log monthly disposable income. * $p < .10$, ** $p < .05$, *** $p < .01$.

Selected and Control. In Selected Replication, the median and average neglect are $\hat{\chi}_i^j = 0.50$ each, while in Nudge the median is $\hat{\chi}_i^j = 0.10$ and the average $\hat{\chi}_i^j = 0.30$.

IV.E. Discussion

In summary, the evidence from the treatments aimed at identifying mechanisms suggests that at least a large part of the reason participants neglect selection in my experiments is that the unobserved signals are not top of mind in the first place, so participants operate with an incorrect mental model and directly use the sample mean to estimate the population mean.

At the same time, these results do not imply that the conceptual or computational difficulty of accounting for selection is unimportant. First, in treatment Nudge, neglect did not disappear despite the fairly strong hint, which suggests that some participants also struggle with the conceptual logic of selection. Second, this experiment was deliberately designed to make overcoming selection both conceptually and computationally reasonably simple, yet doing so is likely much more difficult in real-world applications.

V. Replication

The experiments replace a set of similar experiments, on which an earlier working paper version of this article was based. The earlier experiments followed a very similar logic to the ones described above. Subjects estimated an abstract true state and received computer-generated signals that induced a selection problem of the same kind as above. Although there are a few differences between the earlier experiments and the ones discussed in the main text, perhaps the most important difference is that in the earlier experiments, the true state was based on 15, rather than 6, random draws. Thus, in the earlier experiments, subjects also needed to account for the base rate in processing selected signals. The new design eliminates this additional difficulty. Because the earlier experiments are very similar to the ones reported above, they can be viewed as a replication or robustness exercise. In particular, the earlier experiments also contained versions of treatments Selected, Control, Nudge, Complex, and Simple. Online Appendix G summarizes these earlier experiments and the corresponding results. These experiments also show that (i) subjects neglect selection on average, (ii) the type distribution exhibits a bimodal structure, (iii) an experimental nudge to consider the off-screen signals has a significant effect on beliefs, and (iv) increasing the computational complexity of the decision problem—while holding the difficulty of accounting for selection constant—increases the frequency of neglect.[15]

VI. Discussion and Related Literature

This article has shown that people have a strong average propensity to neglect selection problems when forming beliefs, even when the information-generating process is known and

15. Apart from providing a replication, the earlier experiments also allow for one extension: a study of the responsiveness of subjects' wrong beliefs to observing others holding different beliefs, even though everybody received the same selected information. To investigate this, I implemented experiments that were similar to treatment Selected, except that after subjects had provided their continuous point belief about the true state, they were shown the beliefs of two randomly selected participants from the same experimental session who completed the same task. Then, subjects were provided with an opportunity to revise their beliefs. However, in the data, subjects appear to be very confident in their own way of looking at the problem and largely abstain from revising their beliefs. See Online Appendix G.6 for details.

transparent. A detailed analysis of the mechanisms that give rise to biased belief updating has highlighted the important role of what comes to mind and the resulting mental models. As reflected by the type distribution of neglect, these mental models appear to be binary in nature: subjects either employ a simplistic (and likely automatic) default model of the environment that ignores unobserved data, or they develop an objectively correct representation. An important result of the analysis is that this neglect should not be thought of as an exogenously given neglect parameter that is constant across individuals or even contexts. Rather, the extent to which subjects neglect selection is partly determined by the computational complexity of the decision problem, and the extent to which the decision maker's attention is drawn to the presence of selection.

As discussed in the introduction, the article's approach and results speak to the informal metaphor of a naive intuitive statistician in cognitive psychology (see Fiedler and Juslin 2006 and Juslin, Winman, and Hansson 2007 for overviews; Brenner, Koehler, and Tversky (1996) and Koehler and Mercer (2009) for applications to selection problems).[16] This metaphor and a simple averaging heuristic also characterize much recent experimental economics work on information-processing (Grimm and Mengel 2020; Eyster, Rabin, and Weizsäcker 2018; Graeber 2018; Enke and Zimmermann 2019), including contemporaneous work on endogenous sample selection problems (Araujo, Wang, and Wilson 2018; Charness, Oprea, and Yuksel 2018; Esponda and Vespa 2018; Jin, Luca, and Martin 2018). Indeed, a long line of work on network experiments has documented that a deGroot-style averaging heuristic often describes behavior in complex situations well. The theme that connects these papers is that people appear to have a general tendency to simplify complex information structures by following an averaging heuristic. What sets this article apart from other contributions is (i) the focus on selection problems under a known data-generating process and (ii) a detailed study of the role of incorrect mental models for neglect, including (iii) an exploration of the effect of computational complexity on

16. Work on the availability heuristic (Tversky and Kahneman 1973) is also related in its focus on salient information. However, experimental evidence for the availability heuristic usually involve showing that irrelevant information influences judgment such as in free-form cued recall problems, while in my experiments, relevant information is neglected.

how people form mental models. Thus, the article is close to other work that focuses on why people make mistakes in contingent reasoning. Other such work has highlighted the importance of inferring from simultaneous versus sequential data (Ngangoue and Weizsäcker 2015; Esponda and Vespa 2016) and of uncertainty (Martínez-Marquina, Niederle, and Vespa 2017).

The article results also contribute to an active theory literature that highlights the importance of incorrect mental models. Frequently, researchers motivate incorrect mental models by appealing to constraints on what is top of mind, and this article provided encouraging evidence in this regard. Going forward, a relevant issue for both the theory and the experimental literature will be to identify and describe (i) which incorrect mental models people form and (ii) how these depend on contextual features that are irrelevant under traditional theories, such as complexity, salience, and environmental cues that activate different memory traces.

HARVARD UNIVERSITY AND NATIONAL BUREAU OF ECONOMIC RESEARCH

SUPPLEMENTARY MATERIAL

An Online Appendix for this article can be found at *The Quarterly Journal of Economics* online. Data and code replicating tables and figures in this article can be found in Enke (2020), in the Harvard Dataverse, doi: 10.7910/DVN/1YYUN3.

REFERENCES

Araujo, Felipe A., Stephanie W. Wang, and Alistair J. Wilson, "The Times They are A-Changing: Dynamic Adverse Selection in the Laboratory," University of Pittsburgh Working paper, 2018.

Azrieli, Yaron, Christopher P. Chambers, and Paul J. Healy, "Incentives in Experiments: A Theoretical Analysis," *Journal of Political Economy*, 126 (2018), 1472–1503.

Benjamin, Daniel J., "Errors in Probabilistic Reasoning and Judgmental Biases," in *Handbook of Behavioral Economics: Applications and Foundations 1*, Vol. 2, pp. 69–186 (Elsevier, 2019).

Bock, Olaf, Ingmar Baetge, and Andreas Nicklisch, "Hroot: Hamburg Registration and Organization Online Tool," *European Economic Review*, 71 (2014), 117–120.

Bohren, J. Aislinn, and Daniel Hauser, "Bounded Rationality and Learning: A Framework and a Robustness Result," CEPR Discussion Paper no. 12036, 2017.

Brenner, Lyle A., Derek J. Koehler, and Amos Tversky, "On the Evaluation of One-Sided Evidence," *Journal of Behavioral Decision Making*, 9 (1996), 59–70.

Charness, Gary, Ryan Oprea, and Sevgi Yuksel, "How Do People Choose between Biased Information Sources? Evidence from a Laboratory Experiment," Working paper, 2018.
Costa-Gomes, Miguel A., and Vincent P. Crawford, "Cognition and Behavior in Two-Person Guessing Games: An Experimental Study," *American Economic Review*, 96 (2006), 1737–1768.
Enke, Benjamin, "Replication Data For: 'What You See Is All There Is'," (2020), Harvard Dataverse, doi: 10.7910/DVN/1YYUN3.
Enke, Benjamin, and Florian Zimmermann, "Correlation Neglect in Belief Formation," *Review of Economic Studies*, 86 (2019), 313–332.
Esponda, Ignacio, and Demian Pouzo, "Berk–Nash Equilibrium: A Framework for Modeling Agents with Misspecified Models," *Econometrica*, 84 (2016), 1093–1130.
Esponda, Ignacio, and Emanuel Vespa, "Hypothetical Thinking: Revisiting Classic Anomalies in the Laboratory," UC Santa Barbara Working paper, 2016.
———, "Endogenous Sample Selection: A Laboratory Study," *Quantitative Economics*, 9 (2018), 183–216.
Eyster, Erik, and Matthew Rabin, "Naïve Herding in Rich-Information Settings," *American Economic Journal: Microeconomics*, 2 (2010), 221–243.
Eyster, Erik, Matthew Rabin, and Georg Weizsäcker, "An Experiment on Social Mislearning," CRC TRR 190 Rationality and Competition Working paper, 2018.
Fiedler, Klaus, and Peter Juslin, *Information Sampling and Adaptive Cognition* (Cambridge: Cambridge University Press, 2006).
Fischbacher, Urs, "z-Tree: Zurich Toolbox for Ready-Made Economic Experiments," *Experimental Economics*, 10 (2007), 171–178.
Fragiadakis, Daniel E., Daniel T. Knoepfle, and Muriel Niederle, "Who Is Strategic?," Stanford University Working paper, 2016.
Fudenberg, Drew, "Advancing beyond 'Advances in Behavioral Economics'," *Journal of Economic Literature*, 44 (2006), 694–711.
Gabaix, Xavier, "A Sparsity-Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 129 (2014), 1661–1710.
Gagnon-Bartsch, Tristan, Matthew Rabin, and Joshua Schwartzstein, "Channeled Attention and Stable Errors," Harvard University Working paper, 2018.
Gennaioli, Nicola, and Andrei Shleifer, "What Comes to Mind," *Quarterly Journal of Economics*, 125 (2010), 1399–1433.
Graeber, Thomas, "Inattentive Inference," Harvard University Working paper, 2018.
Grimm, Veronika, and Friederike Mengel, "Experiments on Belief Formation in Networks," *Journal of the European Economic Association*, 18 (2020), 49–82.
Han, Bing, and David Hirshleifer, "Visibility Bias in the Transmission of Consumption Norms and Undersaving," UC Irvine Working paper, 2015.
Heidhues, Paul, Botond Kőszegi, and Philipp Strack, "Unrealistic Expectations and Misguided Learning," *Econometrica*, 86 (2018), 1159–1214.
Jackson, Matthew O., "The Friendship Paradox and Systematic Biases in Perceptions and Social Norms," *Journal of Political Economy*, 127 (2019), 777–818.
Jehiel, Philippe, "Analogy-Based Expectation Equilibrium," *Journal of Economic Theory*, 123 (2005), 81–104.
———, "Investment Strategy and Selection Bias: An Equilibrium Perspective on Overoptimism," *American Economic Review*, 108 (2018), 1582–1597.
Jin, Ginger, Mike Luca, and Daniel Martin, "Is No News Perceived as Good News? An Experimental Investigation of Information Disclosure," Harvard Business School Working paper, 2018.
Juslin, Peter, Anders Winman, and Patrik Hansson, "The Naive Intuitive Statistician: A Naive Sampling Model of Intuitive Confidence Intervals," *Psychological Review*, 114 (2007), 678–703.
Kahneman, Daniel, *Thinking, Fast and Slow* (New York: Macmillan, 2011).

Kahneman, Daniel, and Dale T. Miller, "Norm Theory: Comparing Reality to its Alternatives," *Psychological Review*, 93 (1986), 136–153.
Koehler, Jonathan J., and Molly Mercer, "Selection Neglect in Mutual Fund Advertisements," *Management Science*, 55 (2009), 1107–1121.
Levy, Gilat, and Ronny Razin, "The Coevolution of Segregation, Polarized Beliefs, and Discrimination: The Case of Private versus State Education," *American Economic Journal: Microeconomics*, 9 (2017), 141–170.
Manski, Charles F., "Measuring Expectations," *Econometrica*, 72 (2004), 1329–1376.
Martínez-Marquina, Alejandro, Muriel Niederle, and Emanuel Vespa, "Probabilistic States versus Multiple Certainties: The Obstacle of Uncertainty in Contingent Reasoning," NBER Working Paper no. 24030, 2017.
Ngangoue, Kathleen, and Georg Weizsäcker, "Learning from Unrealized versus Realized Prices," DIW Berlin Working paper, 2015.
Schwartzstein, Joshua, "Selective Attention and Learning," *Journal of the European Economic Association*, 12 (2014), 1423–1452.
Spiegler, Ran, "Bayesian Networks and Boundedly Rational Expectations," *Quarterly Journal of Economics*, 131 (2016), 1243–1290.
———, "'Data Monkeys': A Procedural Model of Extrapolation from Partial Statistics," *Review of Economic Studies*, 84 (2017), 1818–1841.
Tversky, Amos, and Daniel Kahneman, "Availability: A Heuristic for Judging Frequency and Probability," *Cognitive Psychology*, 5 (1973), 207–232.