Extracting Models From Data Sets: An Experiment*

Guillaume R. Fréchette
NYU

Emanuel Vespa
UCSD

Sevgi Yuksel
NYU

November 19, 2024

Abstract

We experimentally study the types of mental models people form by learning from a set of observations. Specifically, we study the kinds of inferences individuals make regarding the statistical relationships among variables from examining data. Remarkably, participants identify the correct correlation structure 61% of the time, despite having almost no information about the data-generating process. The remaining 39% of the cases correspond to two recurring errors: The first error, observed 13% of the time, involves misidentifying the underlying correlation structure and occurs most often in the presence of confounding variables, rising to 28% in such cases. The second error, the most common overall (observed 26% of the time), involves a failure to identify *any* correlations. On the one hand, the first error results in small losses in the most common occurrences; on the other hand, the second error is associated with randomizing behavior, creates substantial losses, and indicates significant difficulties in learning from the data. Importantly, participants display a high degree of consistency in the types of mistakes they make. Hence, different subjects draw opposing conclusions on what constitutes optimal behavior when presented with identical information and those differences are predictable.

[*] We would like to thank numerous seminar participants for helpful comments. We thank Jay Cizeski, Jimena Galindo, Yiyuan Hu, Jie Liao, and Dingyue Liu for outstanding research assistance. This research was supported by the National Science Foundation under Grants 2315663, 2315664, 2315665, 2437062. We are responsible for all errors. Fréchette: gf35@nyu.edu. Vespa: evespa@ucsd.edu. Yuksel: sevgiy@gmail.com.

1 Introduction

People must adopt and rely on *models* (sometimes referred to as *narratives*) to make sense of the world, forecast future events, and assess the optimality of potential actions. Although a growing literature in economic theory studies the implications of adopting incorrect mental models, little is known about what type of mental models people are drawn to and how this is affected by the underlying environment.[1] In this paper, we experimentally study the types of mental models people form by learning from a set of observations. Specifically, we study the kinds of inferences individuals make regarding the statistical relationships among variables from examining data.[2]

Examples of decision makers relying on past observations to learn about the relationships between different variables abound. For instance, in the academic job market, our encounters with previous candidates inform our understanding of what observable attributes (e.g., job market paper and reference letters) predict future performance. Similarly, consumers draw on their experiences to understand how a product's reviews, packaging, and other visible attributes correlate with its quality. Non-professional investors look for patterns in past data to predict a company's value based on its observable characteristics. What is common to all these examples is that the decision maker learns from a limited set of observations without using sophisticated data-analysis techniques. In these situations, do people see patterns that aren't there, or do they simplify the task by focusing on a subset of relevant variables? If deviations occur, are they shared across individuals and occur in specific data structures, or do individuals have distinct *types* that repeat the same mistake across different data structures?

Our experiment is designed to isolate the core element that is common to all these examples: a decision maker forms inferences about the statistical relationships between different variables from a limited set of observations. Participants are presented with several different *data sets*. Each data set corresponds to a set of observations about the functioning of a hypothetical machine and consists of data describing the status of three binary variables: two lights and one sound.[3] Participants are given an opportunity to study these data sets, and take notes about them, with the understanding that at a later stage, they will be asked to make predictions about the same machine (predict the

[1] For recent theoretical and empirical contributions, see Esponda & Pouzo (2016), Spiegler (2020), and Hanna, Mullainathan & Schwartzstein (2014). For more references, see subsection 2 on connections to the literature.

[2] Throughout the paper, we use the term *mental model* to denote a description of how different variables statistically relate to each other. In discussing agents adopting these models, we do not imply they are consciously aware of these models; instead, we posit that their decision-making is consistent with the principles underlying these models.

[3] Our design builds on an experimental paradigm in psychology called *blicket machines* that has been widely used to study causal inference in children (Weisberg, Choi & Sobel 2020).

sound using information on the status of the lights), by solely relying on the notes they have taken earlier, without any access to the original data set. Thus, our experiment produces two types of data: participants' notes and their predictions for each machine. The main object of interest is the participants' predictions, which reveal their understanding of the statistical relationships among the variables. We use these relationships to classify the participants with respect to the different mental models that rationalize their predictions. We study the degree to which these models align or systematically deviate from optimal behavior. The notes serve a dual purpose: First, as an experimental tool, they compel participants to engage with the data and form an understanding of it. Second, they offer insights into how participants organize and learn (or fail to learn) from a set of observations.

The following key features of our experimental design are crucial for the interpretation of our results. First, our implementation is free of context. Outside of the laboratory, contextual cues may affect the inferences people make. However, our focus here is more fundamental: Examining how an agent's mental model is shaped solely by the patterns in a data set, free from the influence of confounding factors. Second, we give participants no information on the data-generating process; that is, we deliberately do not present participants with a set of possible models. Hence, there are no restrictions on the models that participants might consider. Our aim is to shed light on how people learn from a set of observations when they are not directed or suggested to *look* at the data in a specific way. Third, each participant in our experiment encounters multiple data sets, which cover a range of statistical correlations between the variables. These vary both what variables are correlated with one another and the strength of those correlations. This framework enables us to assess the extent to which specific types of mistakes are a function of the environment (aspects of the correlation structure) or an inherent characteristic of the participants (a person-specific trait).

Given the abstract environment and the fact that subjects are provided with very minimal information about how the data are generated, one could expect that participants would fail at identifying the correct correlations. However, our first result shows that, on average, participants do use the information available to them reasonably effectively, accurately identifying the underlying relationships among variables. In other words, average predictions move across data sets in line with what the underlying data generating process suggests. Indeed, participants achieve an average prediction accuracy of 73%, substantially higher than the random benchmark (50%). On the other hand, they are also below the optimal benchmark of 83%, indicating the presence of mistakes. In our setting, optimal behavior must be a deterministic rule (a mapping from lights to sound). Indeed, 78% of predictions are consistent with the use of a deterministic prediction rule—and 61%

correspond to the correct rule.

Overall, there are 39% of cases where our subjects fail to correctly identify the underlying correlation structure. As highlighted in the literature, several types of mistakes are possible. For instance, Hanna, Mullainathan & Schwartzstein (2014) report that experienced Indonesian seaweed farmers remain below the production frontier because they ignore a crucial variable by failing to recognize the impact of pod size, which is a key input in their production process. The informational redundancies problem discussed in Akerlof & Shiller (2010) provides an example for a different of mistake. A realtor reads a newspaper article arguing housing prices tend to increase, and later has a conversation with a colleague who, based on the same article, repeats the information. A naive realtor might think her friend's opinion provides additional information. More generally, this mistake is associated with a misunderstanding or neglect of confounding variables. In our environment, we classify mistakes into two categories. The first involves a failure to condition on any variable at the prediction stage, which we refer to as a *non-conditioning error*: for instance, simply predicting the same outcome irrespective of the data would fall in this category. The second mistake involves conditioning on observable variables but failing to do so in the optimal manner, namely conditioning on an irrelevant variable or conditioning on some, but not all relevant variables. This mistake is labeled a *misalignment error*.

Non-conditioning errors are the most common and are present at a fairly constant rate across: in almost all data sets they describe 21% to 30% of participants. The literature on narratives often assumes that people will be drawn to best-fitting models. Our results suggest that people often miss important statistical relationships in the data and fail to formulate best-fitting models on their own. Misalignment errors, on the other hand, occur more frequently in data sets containing spurious correlations between the variables (observed 26% of the time). To be explicit, in our implementation, we include data sets representing hypothetical machines where the blue light is correlated with the red light, and the red light, in turn, is correlated with the sound. Thus, the blue light is also correlated with the sound, but this is true only when the red light is not conditioned on. Someone who understands the working of the machine would know that given the red light, the blue light carries no further information about the sound. Thus, to predict the sound they should only make use of the status of the red light, but not the blue light. The most common manifestation of misalignment errors in our data involves conditioning on the blue light in data sets of this type. In other words, this result is consistent with a failure to understand conditional independence. Keeping the underlying correlation structure constant, weakening the correlation between variables generally increases non-conditioning errors. The only exception to this are the data sets described above where the blue light is spuriously correlated with the sound, where we observe a rise in misalignment errors instead.

Further analysis reveals that these two types of mistakes point to an important underlying distinction. The misalignment errors observed are well calibrated in the sense that they generate highly accurate predictions for common realizations of observables associated with clear statistical patterns, where deviation from optimal behavior would be most costly. Thus, these mistakes may be seen as attempts to simplify a difficult learning task. The non-conditioning errors, however, cannot be rationalized in this way. Indeed, those are not only cases of non-conditioning, but in addition they involve stochastic predictions. As such, they cannot be rationalized as subjectively Bayesian behavior. Given our setup, non-conditioning by always predicting the sound can be rationalized by a decision maker using the *best* simple model, but randomizing cannot. In this sense, non-conditioning errors of this type reflect substantial difficulties in learning from the data. At the other extreme, among subjects who predict optimally, many display a rather sophisticated understanding. In particular, if they are presented with the machine described above (the blue light is correlated to the red light, and the red light is correlated to the sound), not only do they ignore the blue light when both lights are visible, if we hide the red light, they start conditioning on the blue light; exploiting its unconditional correlation with the sound.

Finally, we document substantial and persistent heterogeneity across participants. Participants who learn (or fail to learn) effectively in one data set are likely to do so in other data sets. For instance, whereas 10 percent of participants achieve a prediction accuracy within five percentage points of the optimal benchmark in *all* datasets, 20 percent of participants are not able to achieve such a prediction accuracy in *any* data set. Furthermore, their mistakes are of the same nature across data sets and these mistakes are associated with distinct ways of analyzing observations, as revealed in the notes participants take. Hence, subjects often draw opposing conclusions on what constitutes optimal behavior after being provided with identical information. In addition, those conflicting opinions are predictable given the consistency in a subject's behavior across data sets. In other words, our results suggest a mechanism for how sustained disagreement can arise in societies. This channel might be particularly relevant for understanding phenomena in diverse settings such as political polarization or excess trading in financial markets, where individuals often access the same information but interpret it in conflicting ways.

2 Connections to the Literature

A growing literature in economic theory studies how incorrect mental models can influence an agent's beliefs and actions.[4] A central premise of this research is that the degree to which an agent adjusts her beliefs or actions in response to experience, or is influenced by others' arguments, is constrained by her initial, potentially incorrect mental model. These papers provide key insights into the consequences of adopting incorrect mental models, and as we discuss later, an emerging body of empirical research has begun to explore these insights. Our approach in this paper differs from the existing literature in that, rather than studying the consequences of specific misconceptions, we focus on understanding when and how these misconceptions arise in the first place.

Misconceptions—incorrect understanding of one's environment—can arise when people struggle to accurately comprehend the situation they are in. Several recent experimental studies support this idea. For example, participants often ignore described correlations when making decisions (e.g., Eyster & Weizsäcker (2010), Enke & Zimmermann (2019)), behave as if their vote matters even when it is not pivotal (e.g., Esponda & Vespa (2014, 2021)), or act in a second-price auction as if it were a first-price auction (e.g., Cason & Plott (2014); see also Martin & Muñoz-Rodriguez (2019)).[5] These studies suggest misconceptions can arise from challenges in reasoning through the description of an environment.[6] By contrast, this paper focuses on misconceptions that stem solely from analyzing data, when no information about the data-generating process has been provided. In other words, our experiment contributes to this literature by providing insights on what types of mental representations arise when participants rely exclusively on data, without knowledge of the underlying rules or institutional details.

In light of these differences, we provide an overview of the experimental literature on the topic. A common approach in the recent literature is to present participants with data sets and an in-

[4] For recent papers on learning under misspecifications, see Esponda & Pouzo (2016), Fudenberg, Romanyuk & Strack (2017), Bohren & Hauser (2021), and Heidhues, Kőszegi & Strack (2018). Spiegler (2020) provides a review of the approach to causal misconception, utilizing tools from computer science (e.g. Pearl (2009)).

[5] Additional examples include participants not understanding equilibrium effects (Dal Bó, Dal Bó & Eyster (2018)), having difficulties in learning from prices (Ngangoué & Weizsäcker (2021)), adverse selection (Charness & Levin (2009), Martínez-Marquina et al. (2019), Ali et al. (2021)), or difficulties in understanding sample selection (Esponda & Vespa (2018), Araujo, Wang & Wilson (2021), Enke (2020)).

[6] Evidence also suggests providing participants with corrective data or experience may not always be effective in correcting such misconceptions. For instance, Esponda, Vespa & Yuksel (2024) demonstrate many individuals who misunderstand aspects of Bayesian updating fail to learn from informative data, even after substantial exposure. Similar challenges are observed in correcting for sample selection biases (Esponda & Vespa 2018) and in addressing incorrect beliefs (Fudenberg & Vespa 2019).

terpretation (e.g., a causal narrative) that rationalizes the data. This interpretation may either correspond to the true data-generating process or be incorrect and therefore misleading. Regardless of their accuracy, these interpretations may be effective in terms of influencing the agent's understanding of the data. Charles & Kendall (2024) demonstrate, through a series of experiments, how an exogenously provided causal narrative can impact and shape participants' understanding of data. They show that giving the same data set to two groups, but providing each with a different narrative, can lead to different choices, in line with the predictions of Eliaz & Spiegler (2020). Their work also illustrates how some narratives can arise endogenously when participants are asked to provide advice to others. Relatedly, Ambuehl & Thysen (2024) investigate how individuals choose between different causal models. Specifically, participants are *simultaneously* presented with a data set and at least two competing narratives.[7] Their findings reveal behavioral heterogeneity: some participants are drawn to models promising favorable outcomes, others take a more cautious approach, and some evaluate models based on their fit with the data. Finally, Barron & Fries (2024) develop a test of model persuasion within the framework of Schwartzstein & Sunderam (2021), showing narratives are particularly effective at changing beliefs when they align more closely with the data. While these papers primarily focus on the combined influence of narratives and data sets on decision making, our goal is to study how participants learn from a data set when no information about the data-generating process is provided.

In this respect, Kendall & Oprea (2024) is the most closely related work to ours, as it similarly examines how participants make inferences about the data-generating process by analyzing a set of observations. However, unlike our approach, Kendall & Oprea (2024) employ data-generating processes modeled as finite automata, which dynamically connect past outputs and current inputs to future outcomes. In contrast to our findings, their results highlight significant challenges in learning, emphasizing how the underlying structure of the data, such as serially connected observations, can limit learning and inference.

There is also recent work in the other direction: eliciting narratives—causal models—from people about real-world phenomena. For example, Andre et al. (2021) elicit narratives that people and experts use to explain macroeconomic phenomena. A related body of research examines how narratives influence information acquisition. For example, Bursztyn et al. (2023) show that individuals with opposing narratives about the pandemic seek out different opinion programs, even when offered substantial incentives to learn objective facts. The key difference between these studies and

[7] Charles & Kendall (2024) also have participants face multiple narratives, but a second narrative is presented after participants have made decisions knowing the first narrative.

ours is that, in our case, participants are provided only with exogenously supplied datasets to learn about their environment. In other words, we are not investigating how subjects choose the data they use; rather, we provide a dataset and examine how they engage with it.

Finally, a literature in cognitive psychology documents how humans form causal models from an early age (e.g., Weisberg, Choi & Sobel (2020)). As described in Rottman (2017), the study of causal reasoning in adults can be divided into two groups. The first focuses on whether, given data and a set of possible causal models, people can identify the correct model (e.g., Steyvers et al. (2003)). The second provides participants in experiments with a causal model and studies how this knowledge impacts their reasoning in subsequent tasks (e.g., Rottman & Hastie (2014), Fernbach, Darlow & Sloman (2011)). A difference from either approach is that we do not provide participants with any information other than a data set and study whether their choices are consistent with some model. In this sense, our approach is closer to a literature in cognitive psychology that can be broadly described as *associative learning*, which studies how people can form models of the world by studying association of events (for a survey, see Le Pelley et al. (2017)). A classic illustration includes Pavlovian learning, where one event signals another; for instance, clouds in the sky indicate rain may follow. Although many papers in this literature focus on situations with context, evidence suggests distance between events—which can be taken as a measure of how noisy the association is—has an impact on the extent to which people make use the association. As we describe later, our finding on the impact of noise in our abstract environment is consistent with results in this literature.

3 Experimental Design

We first provide an overview of the experimental design. Then, we describe the experiment in further detail from the perspective of what was presented to the participants. Other aspects of the experimental design are described later in this section.

3.1 Task

Each session consists of two parts. In Part 1, participants are presented with 11 different *data sets*, one at a time. After seeing all data sets, participants proceed to Part 2, where they are asked to make predictions for each data set they observed in Part 1. Two important features link Part 1 to Part 2: (i) The Part 1 data set is informative about predictions in Part 2 but is *not available* to the participants during this second part; and (ii) when participants are presented with data sets in Part 1, they can type *notes-to-self* on the computer terminal and these notes *will be available* to them during the relevant prediction tasks in Part 2.

The framing used in the instructions is that each of the 11 data sets is generated by a different *machine*. A machine consists of lights of different colors and can make a sound.[8] Each data set consists of 27 trials, where each trial records an occurrence of the machine's operation. For each trial, the record shows the status of two lights—red and blue—whether they were on or off, and whether the machine made a sound.[9] In the paper, we use variables R, B, and S to refer to the red light, blue light, and the sound, respectively, and denote their status by assigning values 0 or 1 to these variables. The instructions (reproduced in Online Appendix G) are carefully worded to avoid a direct suggestion of a causal relationship between the lights and the sound.[10]

An example of a data set from Part 1 is shown in Figure 1. The 27 trials are presented on the right side of the screen, all at once, where each trial is a row.[11] On the left side of the screen, the participant can take notes. These notes are made available to them in Part 2 when they make predictions. In Part 1, the participant sees 11 screens such as this one, one for each data set.[12]

When participants face Part 1, they also know the prediction task that they will face in Part 2. In general, subjects know they will be asked to make a prediction on whether the machine makes a sound conditional on the status of some of the other variables, namely, the lights. As a reference, we present an example in Figure 2. The participants can see their notes and the status of the red and blue lights. They are asked to guess whether the machine makes a sound. The order of presentation for the machines, the trials within each *data set*, as well as the prediction tasks are randomized across subjects.[13] At the end of the session, one prediction task is selected at random, and if their prediction is correct, they receive $25, in addition to the $10 show-up fee.

[8] As discussed in the literature review, this design is an adaptation of the so called *blicket machines*, a device used in experiments with children (Gopnik et al. 2004) to study causal inference.

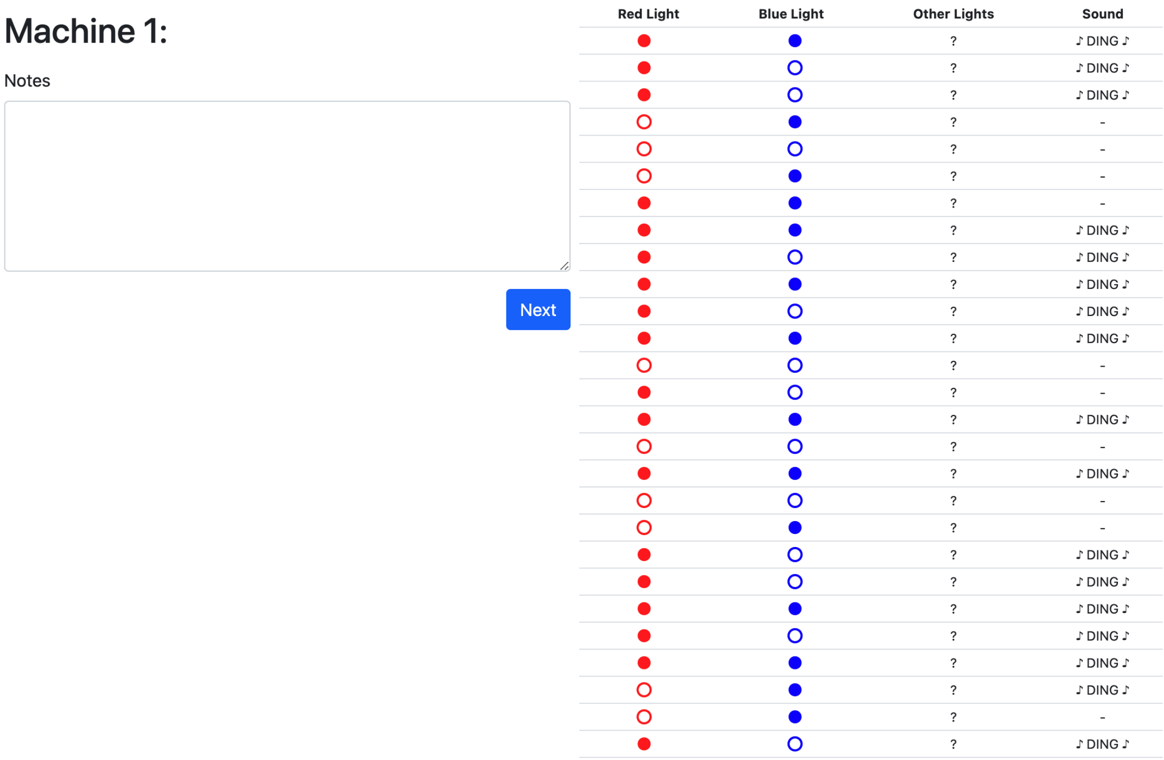
[9] Although the status of only two lights (red and blue) is reported, as a way to motivate the potential probabilistic nature of how different variables (the lights and the sound) are related, participants are told the machine may include other lights, the status of which they will not be informed about.

[10] The instructions include the line "The lights and the sounds may or may not be related to each other."

[11] We used 27 trials because we could fit 27 rows at most on the computer screens used in the laboratory with no need to scroll up or down.

[12] Once participants have taken notes for a data set and they move on to the next one, they cannot return to the previous ones.

[13] The last nine prediction tasks in treatment Unspecified, discussed in section 3.2, are an exception.



Machine 1:

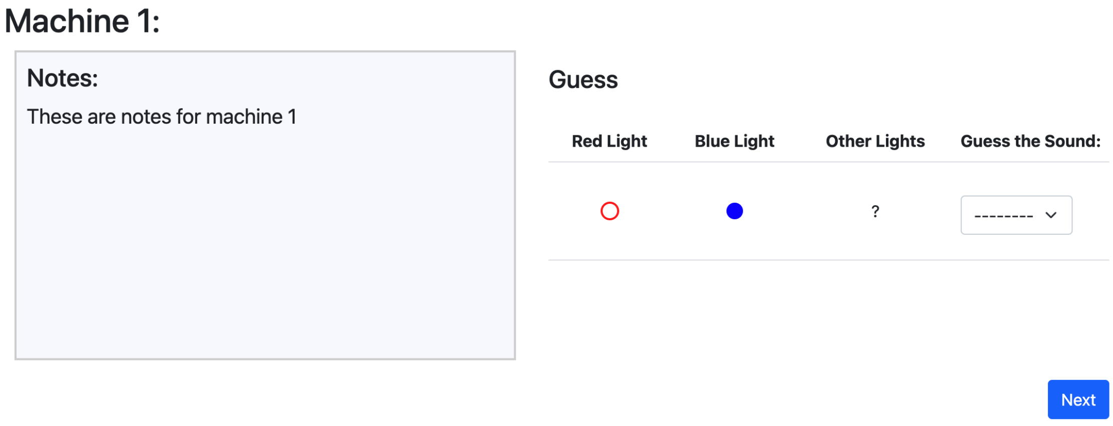
Notes

Next

Red Light	Blue Light	Other Lights	Sound
●	●	?	♪ DING ♪
●	○	?	♪ DING ♪
●	○	?	♪ DING ♪
○	●	?	-
○	○	?	-
○	●	?	-
●	●	?	-
●	●	?	♪ DING ♪
●	○	?	♪ DING ♪
●	●	?	♪ DING ♪
●	○	?	♪ DING ♪
●	●	?	♪ DING ♪
○	○	?	-
●	○	?	-
●	●	?	♪ DING ♪
○	○	?	-
●	●	?	♪ DING ♪
○	○	?	-
○	●	?	-
●	○	?	♪ DING ♪
●	○	?	♪ DING ♪
●	●	?	♪ DING ♪
●	○	?	♪ DING ♪
●	●	?	♪ DING ♪
○	●	?	♪ DING ♪
○	●	?	-
●	○	?	♪ DING ♪

Figure 1: Screenshot of Part 1

Notes: The participants can leave *notes-to-self* after inspecting the data set presented on the right side of the screen. Each trial is a row in the right-hand-side table. A light that is on (off) is represented with a full (hollow) circle. The sound is captured with "Ding," and no sound, with a dash (-).



Machine 1:

Notes:

These are notes for machine 1

Guess

Red Light	Blue Light	Other Lights	Guess the Sound:
○	●	?	--------

Next

Figure 2: Screenshot of Part 2

Notes: The participant can read the *notes-to-self* they wrote in Part 1. In this example, they are provided with the status of the red and blue lights—hollow (full) means light off (on)—and a drop-down menu lets them guess whether the machine makes a sound.

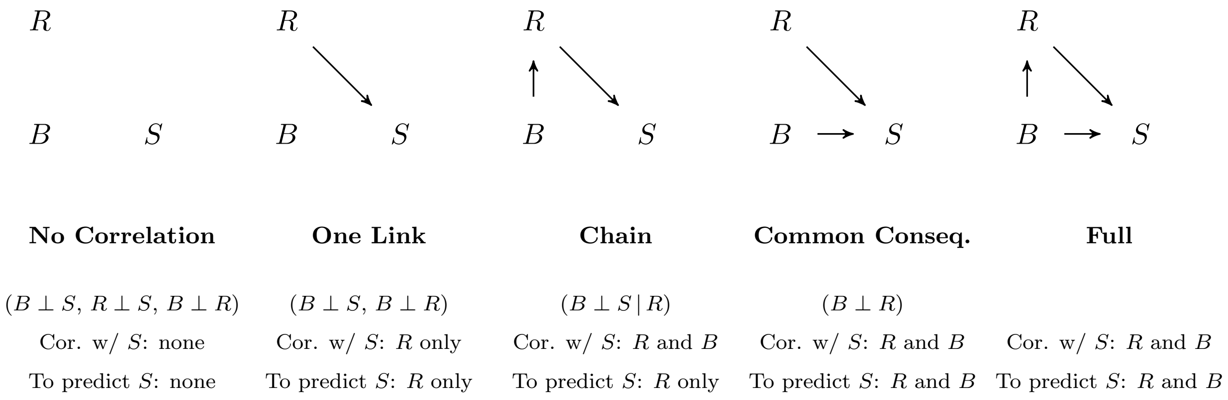


Figure 3: Directed Acyclic Graphs (DAGs)

Notes: R, B, and S are binary variables that capture the status of the red light, blue light, and sound, respectively. The independence conditions implied by the DAG are reported in parentheses. For instance, in No Correlation, all variables are independent from others, and in Chain, B is independent of S conditional on R. Below the independence relations, all variables correlated with S are reported. Finally, the last row reports the variable(s) one would optimally condition on to predict S when both R and B are observed.

3.2 Within-Subjects Treatments: 11 *Data Sets*

In this section, we describe in three stages how we generated the 11 data sets presented to the participants. First, we present a set of directed acyclic graphs (DAGs), which formed the basis for how we created the data sets.[14] Second, we outline the process by which we chose 11 distinct sets of parameters—each referred to as a *parametrization*—associated with these DAGs. Third, we describe how these parametrizations were transformed into data sets. This process generates variation in *data sets* with the aim of understanding how participants' behavior—namely, their ability to extract information from a set of observations—changes with the statistical relationship between variables, the amount of noise, and the complexity of the *data set*.[15]

Selected DAGs

We focus on five DAGs involving R, B, and S. These DAGs and their key characteristics are represented in Figure 3.[16]

In *One Link*, R and S are the only variables that are correlated. As a consequence, the optimal

[14] DAGs are often used to describe causal relationships between variables. Because participants simply face a prediction task, causality does not play a role in our experiment. We use the DAGs as an easy way to summarize possible correlation structures.

[15] More detailed descriptions of each data set are provided in Online Appendix A.

[16] To ease presentation, we fix colors so that, for instance, B is independent from S in One Link. In practice, however, the design varied the meaning of B and R.

prediction for S conditions on R.[17] *Chain* is closely related to *One Link*, with the difference being the additional link between B and R. Here, not only R but also B is correlated with S. Notice, however, that conditional on R, B and S are independent. For this reason, when the status of both lights are known, an optimal prediction for S only conditions on R. Note, however, that in contrast to *One Link*, a prediction for S might optimally condition on B in *Chain* when the status of R is not known. *One Link* and *Chain* therefore provide an interesting comparison, which we come back to in later sections of the paper.

In *Common Consequence*, the outcome S can be interpreted as the common consequence of two variables, R and B. Both R and B are correlated with S, and the optimal prediction rule for S conditions on both variables. In fact, we consider two versions of common consequence, depending on whether R and B jointly generate the effect on S (AND condition), or independently generate the effect (OR condition). *Full* changes the statistical relationship between the variables relative to *Common Consequence (OR)* in a similar manner to how *Chain* compares with *One Link*, that is, by adding a link between B and R. Here, the optimal prediction for S conditions on both lights, the same as with *Common Consequence.*

As a benchmark, we also include a DAG, *No Correlation*, in which all variables are independent, where the optimal prediction for S does not condition on R or B. Note that as we move from *No Correlation* to *Full*, the statistical relationship between the variables becomes increasingly more complex. Focusing on the number of variables that are correlated with S, variation from none (*No Correlation*) to one (*One Link*) and two (*Chain*, *Common Consequence* and *Full*) exists. Focusing on the number of variables that is optimal to condition on to predict S (when status of both lights are observed), variation from none (*No Correlation*) to one (*One Link* and *Chain*) and two (*Common Consequence* and *Full*) exists.

From a DAG to *Parametrization*

There are many different joint probability distributions over the three variables (R, B, and S) consistent with each DAG.[18] We pick 11 *parametrizations* based on the DAGs discussed above: one parameterization for *No Correlation*, and two parameterizations for *One Link*, *Chain*, each version of *Common Consequence* (AND and OR), and *Full.* In DAGs with two parametrizations, we manipulate the strength of the statistical relation between the different variables. In the *Low noise*

[17] To ensure this, we choose parametrizations where $p(S = 1 \mid R = 1) > 0.5$ and $p(S = 1 \mid R = 0) < 0.5$.

[18] For instance, consider *One Link* in Figure 3. The DAG does not specify the likelihood of $B = 1$ or $S = 1$; it simply imposes that these events are independent. Similarly, how exactly R is correlated with S is also not pinned down beyond the requirement that these variables are not independent.

parameterization, if an arrow goes from one variable to another, the probability that the realization of the latter one matches the former is equal to 90 percent. In the *High noise* parameterization, the corresponding probability is only 80 percent.[19]

Online Appendix A covers the specific way in which we chose the 11 parametrizations. Here, we describe the criteria that underlie these choices. First, the probability that $S = 1$ is set to approximately 0.62 in all 11 cases. This implies that a prediction rule for S that does not condition on the lights can achieve an accuracy of approximately only 62 percent in all cases, which is better than random, while still generating sufficient incentives to use the observables (the status of the lights) to predict the sound. Our second criterion aims to improve identifiability of whether a participant is using the optimal prediction rule for a given parametrization. As a reference, consider *One Link*: in both the low- and high-noise conditions, if the participant is provided with the status of the lights and is asked to predict the sound, the optimal strategy is to guess the machine will make a sound only when the red light is on—a *deterministic prediction rule* we refer to as *G w/ R*, for "Guess when $R = 1$." Such a rule achieves a prediction accuracy of 90 or 80 percent, depending on the noise condition (as shown in Table 6 of Online Appendix B). Using an alternative prediction rule such as guessing the sound is on whenever the red light or the blue lights are on (referred to as *G w/ R or B*) lowers prediction accuracy to 76 and 70 percent in the low- and high-noise conditions, respectively. We chose the parameters for each case (subject to the constraints stated above) to increase as much as possible the prediction accuracy cost associated with using a suboptimal prediction rule.

To summarize, we chose the 11 parameterizations to (i) keep the likelihood of $S = 1$ around 62 percent; (ii) satisfy the different noise conditions; and (iii) increase the cost of using a suboptimal prediction rule.

From Parametrization to *Data Set*

Parameterizing the DAGs moves us a step closer to creating the data sets that were presented to the participants. Specifically, for each of the 11 parameterizations, we can compute the joint distribution over realizations of the lights and the sound. If participants were to see a large data set, the frequency of observing different trials in this data set (corresponding to different light and sound configurations) would closely match this distribution. However, to make the task of learning

[19] When two arrows point toward the same variable (as in Common Consequence and Full DAGs), we assume errors are independent in the OR condition, but correlated in the AND conditions.

from data manageable for subjects, we limited the size of the data set shown to the participants.[20] For each of the 11 parameterizations, we created a 27-trial data set where empirical frequencies over different trials are closest to the true probability distribution.[21] All participants are presented with the same 11 data sets. Thus, these data sets create the main source of within-subjects variation in our experimental design.

Between-Subjects Treatments

In addition to the within-subject variation in terms of the *data sets* that participants face, we conducted two manipulations that we implemented across subjects. In the *Specified Prediction* treatment, we informed participants that Part 2 predictions consist of guessing whether the machine will make a sound after they are shown the status of the red and blue lights. We have two implementations of this treatment. In the first implementation, we constrained notes to 280 characters. In the second implementation, we constrained notes to 75 characters.[22]

In the *Unspecified Prediction* treatment, participants were told that Part 2 predictions may consist of (1) predicting the sound after learning the status of both, one, or none of the lights; or (2) predicting one light after observing either the other light, the sound, both, or none. Notes were constrained to 75 characters. This treatment allows us to study to what extent participants can adjust their predictions when they have only partial or no information on the status of the lights.

Part 2 Predictions

In the Specified Prediction treatment, participants faced 27 prediction rounds associated with each machine in Part 2.[23] For each machine, prediction rounds were presented one by one and in random

[20] The design deliberately avoids providing subjects with “data summaries” that report relative frequency of different trials, because such an approach would already suggest a specific way of learning from data. See Esponda et al. (2024) for how providing summarized data can significantly alter learning outcomes.

[21] Randomly drawing 27 trials independently for each subject using the true probability distribution would have created significant variation across subjects in the data sets presented to them. Although this approach would have generated rich data in other ways, it would have made it difficult (i) to control incentives and (ii) to identify whether subjects are using the optimal prediction rule.

[22] With the initial character limit of 280 (which used to be the limit on Twitter), a non-negligible proportion of subjects were transcribing the entire data set directly in the notes. To identify the extent to which results depended on such behavior, we conducted a second implementation that constrains note length for the median participant from the first implementation.

[23] Participants were asked to predict the 27 trials that they observed in the Part 1 *data set*. However, this was not revealed to the participants. In the instructions, they were only told that in Part 2 they would make several predictions for each machine.

order. Observing many predictions for each machine is useful to identify the prediction rules being used, because we can observe participants making predictions for the same light configurations a number of times.

In the Unspecified Prediction treatment, just as in the Specified Prediction treatment, participants made the same 27 predictions. In addition, they made nine more predictions, which were presented after the initial 27. As in earlier rounds, these predictions were also about the sound, but participants were provided with partial or no information on the status of the lights.[24]

Implementation Details

We conducted four sessions of each treatment at the EconLab in UC San Diego, where participants are students at the university. The instructions, which are available in Online Appendix G, describe the Part 2 prediction task but do not specify the number of predictions that participants will make per machine. After reading instructions, participants could move at their own pace but were informed they would not be able to leave early. A total of 88, 72, and 70 participants are in Specified Prediction with the 280-character note limit, Specified Prediction with the 75-character note limit, and Unspecified Prediction with the 75-character note limit, respectively.

3.3 Discussion of the Design

We designed the experiment with the following goals in mind.

First, we adopt a context-free implementation by framing datasets as corresponding to abstract machines. Although this approach abstracts away from other important forces that could influence an agent's understanding of their environment, it allows us to study how an agent's mental model is shaped by the patterns they identify in observations in the absence of any confounds. Driven by similar considerations, we do not focus on dynamics of learning. For this purpose, we provide data sets in their entirety all once to our participants. Many interesting questions could be asked about how context or past experiences within the same environment impact learning, which we hope can be tackled in future work.

Second, we provide no information about the data-generating process. Specifically, participants are not given a set of models or a prior over these models. This unstructured design approach

[24] Specifically, there were four rounds where only the status of the red lights was revealed (two with $R = 1$ and two with $R = 0$), four rounds where only the status of the blue lights was revealed (two with $B = 1$ and two with $B = 0$) and one round where no information on the lights was provided.

places no restrictions on the range of possible models participants may consider. For example, a participant might speculate about whether serial correlation exists between different observations or whether one data set informs another. Additionally, because each data set is presented all at once, no information is available on the timing of events (e.g., the sequence of lights and sound), which is known to affect causal inference (see, e.g., Bramley et al. (2018)). Our aim is to shed light on how people learn from a set of observations when they are not directed or suggested to *look* at the data in a specific way. Motivated by similar reasons, we never ask participants directly about their "models," as this line of questioning might influence how they learn from their observations. Instead, we study how people make predictions, an exercise that has a natural counterpart in many real-world applications. These predictions "reveal" participants' mental models, that is, how predictions condition on different observable variables reflect the inferences they made regarding the statistical relationship between variables.

Finally, each participant in our experiment encounters 11 data sets, which cover a range of possibilities with respect to the statistical correlations between the variables. This approach enables us to examine whether a participant's ability (or inability) to adopt optimal models remains consistent across different environments, or if it is shaped by the features of the environment.

4 Results

4.1 Do Participants Leverage Information from Data Sets to Make Predictions?

Prediction Accuracy and Optimality

Our participants face an abstract environment that may be challenging to process. Hence, a natural starting point is to evaluate the extent to which they can use information in the data sets to make predictions. Specifically, we evaluate the accuracy and optimality of participants' predictions. Prediction accuracy is defined as the probability with which a participant's guess is correct. Optimality of a prediction, on the other hand, is coded as either 0 or 1, capturing whether the prediction maximizes accuracy.[25] Figure 4 plots the cumulative distribution of these measures computed at the participant level for different data sets separated into three categories: the No Correlation data set, the High Noise data sets, and the Low Noise data sets. The vertical lines in

[25] For each data set, we compute the probability that the machine makes a sound conditional on the lights' status: $Pr(S = 1|B, R)$. The accuracy of a prediction is $Pr(S = 1|B, R)$ or $Pr(S = 0|B, R)$ depending on the direction of the prediction. The prediction is optimal if it maximizes accuracy.

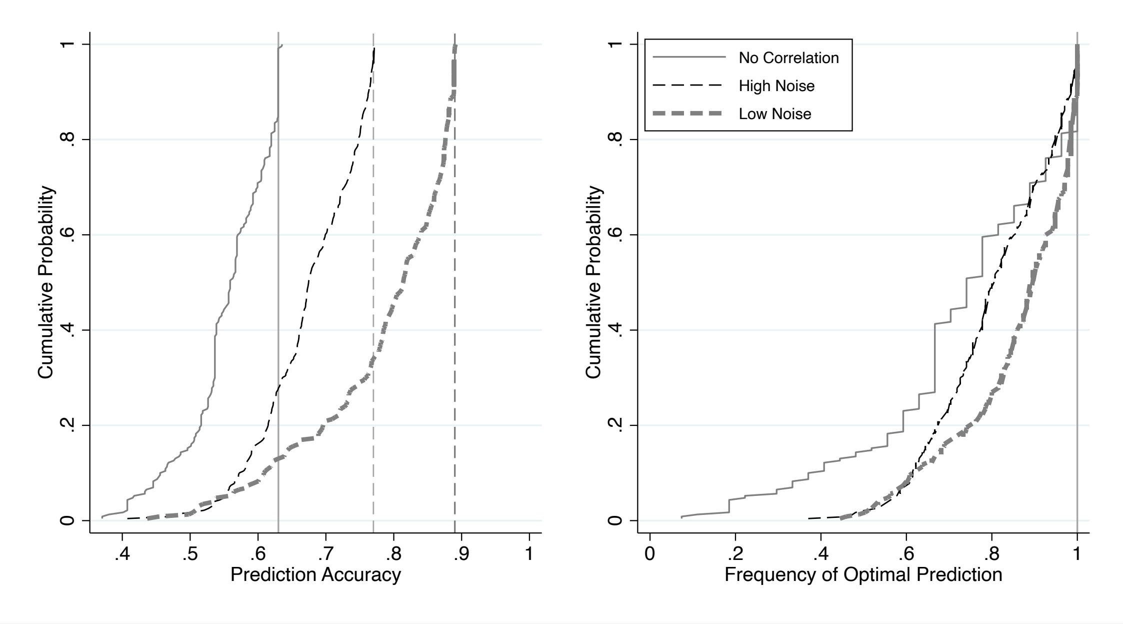


Figure 4: Distributions of Prediction Accuracy and Optimality

Notes: In the left panel, each observation corresponds to the expected prediction accuracy of a subject in the data sets corresponding to the specific category, and the vertical lines denote the prediction accuracy for an agent who guesses optimally in the three respective categories. In the right panel, each observation corresponds to the frequency with which a subject's guesses were optimal given the available information in the data sets corresponding to the specific category, and hence, the best achievable value is 1 for all three categories.

the left panel depict the level of prediction accuracy that could have been achieved by an agent who always makes optimal predictions in the three corresponding categories. As expected, this benchmark depends on how informative R and B are with respect to S, and hence is lower in the High Noise data sets than in the Low Noise data sets, but lowest in the No Correlation data set where S is independent of R and B. In the right panel focusing on prediction optimality, by definition, the best achievable value is 1 for all three categories.

A first observation when focusing on the left panel is that in all three categories, the vast majority of participants make predictions that are more accurate than random, which corresponds to a prediction accuracy of 0.5. This finding shows participants are able to leverage information in the data set to improve their prediction accuracy. In No Correlation, 15 percent of participants do worse than random. Because in this data set lights carry no information, performing better than random indicates most participants do use information on $Pr(S = 1)$, the only feature of the data set with value for predictions. In fact, 19 percent of participants *always* make optimal predictions in this data set.

If participants ignore the lights and focus only on $Pr(S = 1)$, they would achieve similar accu-

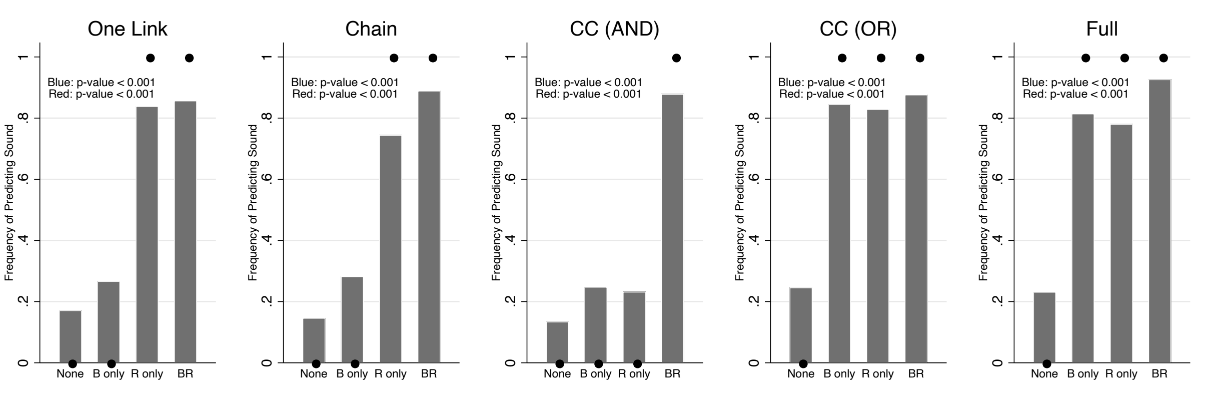


Figure 5: Predictions by Light Configuration

Notes: *None* refers to trials where both the red and blue lights are off; *B only* (*R only*) refers to trials where only the blue (red) light is on; *BR* refers to trials where both lights are on. Black dots denote optimal behavior. Reported p-values refer to F-tests on whether prediction frequency changes with the status of the blue and red lights. For blue (red), we test jointly whether the prediction frequency changes from *None* to *B only* and from *R only* to *BR* (from *None* to *R only* and from *B only* to *BR*).

racy levels in all data sets, as $Pr(S = 1)$ is approximately 62 percent, by construction, in all cases. Consequently, changes in prediction accuracy when comparing Low or High Noise data sets with No Correlation indicate participants are also using information on how the lights correlate with the sound to make their predictions in these data sets. Indeed, we observe that the distributions of prediction accuracy sharply improves relative to No Correlation. In fact, we see a clear ordering depending on the amount of information in the data set. Prediction accuracy is relatively worse in No Correlation, followed by High Noise, followed by Low Noise. Interestingly, as seen in the right panel of Figure 4, this ordering is preserved with respect to the optimality of predictions.[26,27] Overall, these patterns suggest that most subjects are engaged, are extracting payoff-relevant information from data sets, and perform better in data sets with stronger correlation between variables.

Figures 11 and 12 in Online Appendix C reproduce the information in Figure 4 separately for each of the three between-subjects treatments. These figures reveal the variation across these treatments—limiting the number of characters in the note-to-self and having a broader prediction task—have no significant impact on behavior with respect to accuracy or optimality.[28] For this reason, in presenting results, we pool data from these treatments.

[26] Equality of the distributions can be rejected by a Kolmogorov-Smirnov test (p value < 0.001 in all pairwise comparisons).

[27] Note, the share of participants whose predictions are optimal more than 80 percent of the time increases across these categories from 40 percent in No Correlation to 52 percent in High Noise, to 74 percent in Low Noise.

[28] Tables 9 and 10 in Online Appendix E also document the degree to which these treatments impact the likelihood of different types of mistakes.

Furthermore, having established that participants are able to extract information from data sets, in the remainder of the paper, we focus our analysis on the 10 data sets (other than No Correlation) in which the observables (configuration of the lights) provide some information on the status of the sound.

Predictions as a Function of Lights

The analysis so far does not yet demonstrate how predictions condition on the different light configurations and how this depends on the statistical relationships between variables in each data set. We first aggregate over the Low and High Noise conditions to describe how predictions vary with the different light configurations in each DAG.[29,30]

Each graph of Figure 5 presents the aggregate frequency of predicting $S = 1$ for each possible light configuration (R, B).[31] Black dots denote optimal behavior for each case. As can be seen in all cases, a clear pattern in the direction of optimality always exists. For example, in One Link and Chain, the likelihood of predicting the sound to be on changes by 50 percentage points depending on the status of the red light, but the status of the blue light has a much smaller impact. In Common Consequence (both OR and AND) and Full, as predicted, the status of the blue light also has a large impact on predictions.

So far, patterns provide aggregate-level evidence that in our environment, many participants can extract information and do condition their predictions on the variables they observe (i.e., the status of the lights). In fact, the conditioning varies with the structure of the data set in the direction of optimal behavior. However, Figure 4 also highlights that many subjects are are far from optimal. At the aggregate level, under Low (High) Noise, 19% (29%) of subjects in our sample achieve a prediction accuracy closer to the random benchmark of 50% than to the optimal benchmark. At the DAG level (as seen in Figure 5), predictions deviate from optimality in revealing ways. For example, in Chain, when only the blue light is on, 28 percent of the participants suboptimally predict the sound to be on, suggesting predictions vary (suboptimally) with the status of the blue light. Such mistakes are not unique to Chain. Regardless of the structure of data set, in all cases, in at least one light configuration, predictions are suboptimal more than 20 percent of the time.

Result 1. *Participants, on average, condition their guesses on the observables. Their predictions*

[29] Technically, Common Consequence [AND] and [OR] are not distinct DAGs, but they differ in terms of the optimal prediction rule for S. Thus, we separate out these cases when presenting results at the DAG level.

[30] Figure 13 in Online Appendix C decomposes Figure 5 by noise levels.

[31] Reported p-values in the top-left corner of each graph report results from F-tests on whether the frequency of guessing the sound changes with the status of the blue and red lights (see notes to Figure 5 for details).

for each data set vary in the direction of optimal behavior. However, in all data sets, predictions remain far from optimal for many participants.

The analysis at the aggregate level shows participants are learning from data sets, namely, they are able to form mental models that allow them to condition their predictions on the observable variables, and that such models change with the statistical structure of the data sets. But such analysis is limited because we do not know to what extent, for example, these patterns are driven by a small or large subset of participants. Figure 4, for instance, highlights substantial heterogeneity across participants.

4.2 Can Participants Be Rationalized as Using Deterministic Prediction Rules?

In this section, we examine behavior at the individual level to study the extent to which participants are able to use the optimal prediction rule in each environment. This approach also allows us to identify the kinds of mistakes participants are prone to make in each environment. To study these questions, we type participants as using different prediction rules. A prediction rule is a mapping from R and B to S; namely, it specifies how the prediction of the sound depends on the status of the lights.[32] As described in Section 3 (and listed in Table 5 in Online Appendix B) there are 16 deterministic prediction rules.[33] In addition to these rules, we also consider a stochastic prediction rule that predicts $S = 1$ with some probability p (to be estimated) that is independent of the status of the lights.

We type participants on the data set level using a two-step procedure: (1) We estimate the distribution of prediction rules that are used at the population level; and (2) for each participant, given their guesses, we compute the posterior likelihood of following each prediction rule, using population-level estimates as a prior. We classify participants as using a prediction rule by identifying the highest posterior.

Here, we provide an outline of this typing methodology and refer the reader to Online Appendix D for further details.[34] We use a finite mixture model to estimate the distribution of prediction rules

[32] In the analysis presented in this section, we do not consider prediction rules that condition on the order of events in the prediction task or the data set. The computer interface was designed to minimize the salience or attractiveness of such rules. We also do not find compelling evidence supporting the use of such rules (see Online Appendix E).

[33] These rules differ on whether and how predictions change with R and B. For instance, *G All* is a rule that predicts $S = 1$ for all light configurations and *G w/ R* is a rule that predicts $S = 1$ only when $R = 1$.

[34] This technique was recently used in Aoyagi, Fréchette & Yuksel (2024) to classify subjects into different repeated-game strategies. Simulations (reported in Online Appendix D) demonstrate type shares can be reliably recovered for

used. The method specifies a set of candidate prediction rules (which corresponds to the 17 rules discussed above) and then estimates their prevalence in the population allowing for the possibility of implementation errors. Formally, for each data set, we use the $27 \times 230 = 610$ predictions to estimate 18 parameters: the probability distribution over the the set of rules, implementation error for the rules, and the probability of predicting the sound for the stochastic rule. Then, we use the mixture-model estimates as a prior and compute the Bayesian posterior that a participant is using each of the candidate rules given the set of predictions they make. Each participant is associated with the rule that has the highest likelihood according to this posterior.

This simple typing method successfully reflects the differences in how participants make predictions in our experiment. Using a *participant-data set* as a unit of observation, 78 percent of observations are typed as matching a deterministic rule, although it is worth reiterating that we allow such rules to be implemented with some implementation errors. In 46 percent of these cases, behavior *perfectly* matches the rule; that is, predictions coincide with the rule for *all* 27 rounds. Overall, among *all* observations typed to correspond to a deterministic rule, predictions match the rule 94 percent of the time. Focusing on *participant-data set* observations that are typed as corresponding to a stochastic rule, the overall frequency with which the sound is predicted is 60 percent. Such behavior could possibly be driven by probability matching behavior (i.e., when the predicted frequency of sound matches the observed frequency of sound), which we discuss further in section 4.4.[35]

The impact of noise

A more disaggregated summary of the results is presented in Figure 6, which separates data sets by the noise condition (Low vs. High).[36] Aggregating across Low Noise data sets, we find only 18 percent of observations are typed as not corresponding to a deterministic rule. Meanwhile, of those that are typed as consistent with a deterministic rule, the vast majority are typed as using the optimal rule. In fact, close to two-thirds of the observations correspond to the optimal rule

all data sets used in our experiment by following this estimation procedure. Alternative methods for typing subjects—such as typing participants based on the rule their predictions are most consistent with—generates qualitatively very similar results.

[35] Indeed, modal behavior perfectly aligns with probability matching. Overall, in 44 percent of the cases, the frequency with which participants predict the sound to be on is within five percentage points of the true frequency in the data set. See Figure 16 in Online Appendix E for more information on the distribution.

[36] Table 8 in Online Appendix C reports for each data set the share of observations typed as corresponding to the most popular deterministic rules. In addition, the table also includes information on the share of observations that are not classified as consistent with a deterministic rule.

in low-noise data sets. Of those, about half of them perfectly match the optimal rule (in all 27 rounds). Qualitatively, these findings are replicated with high noise. Two-thirds of observations are classified as consistent with a deterministic rule and the optimal rule is the modal rule that is most consistent with predictions. However, the comparison of low-noise to high-noise data sets reveals a sharp decline in (i) the share classified as corresponding to a deterministic rule, (ii) the share classified as the optimal rule, and the share perfectly consistent with the optimal rule.

Why is the change in the share corresponding to deterministic rules noteworthy? The use of a deterministic prediction rule, optimal or not, indicates a participant can extract statistical patterns—whether correct or incorrect—from the data set, which they then utilize for making predictions. As seen in Figure 6, a significant share of observations (18 and 27 percent, respectively, with low and high noise) are classified as corresponding to stochastic behavior, which does not condition predictions on light configurations. This observation suggests limitations on the extent to which people form mental models by studying data sets, and indicates their ability to do so varies with the strength of the statistical relationships observed in the data sets.

One implication of these results is that participants are more likely to display variation in their predictions in high-noise data sets, which can be captured in a few ways. First, the variance (across subjects) in the overall frequency of predicting the sound increases by 42 percentage points in high-noise relative to low-noise data sets. Second, given a light configuration, the likelihood that any two participants disagree on their prediction (make opposing guesses on the sound) increases by 28 percentage points in high-noise relative to low-noise data sets. Furthermore, disagreement increases not only across participants but also across different predictions by the same participant. That is, the likelihood that two predictions by the same participant for the same light configuration is in disagreement increases by 50 percent in high-noise relative to low-noise data sets.[37] Taken together, these results display behavior to be more variable (both within and across subjects) in high-noise data sets.

Result 2. *The majority of guesses are consistent with the use of a deterministic rule. The optimal rule is the modal prediction rule in low and high-noise data sets. High noise decreases the likelihood that predictions are consistent with the optimal rule and increases variation and disagreement in predictions.*

A sizeable share of participants are classified as not using optimal prediction rules. This finding presents new questions: What types of deviations from optimal behavior do we observe? Are these

[37] Figure 14 in Online Appendix C provides further evidence on these observations.

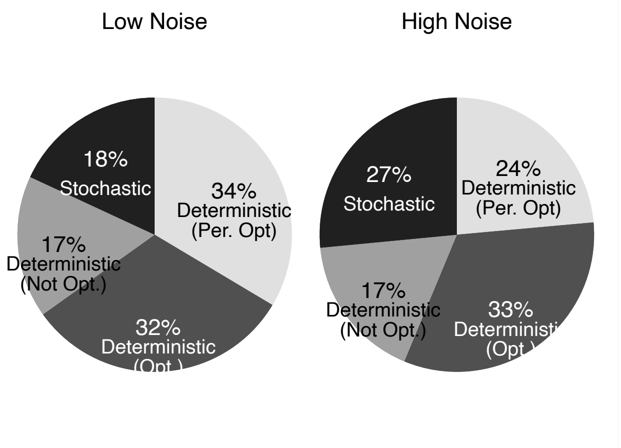


Figure 6: Share of Subjects Using a Deterministic and Optimal Rule

Notes: *Deterministic* (*Stochastic*) refers to whether the subject is typed as (not) using a deterministic rule; *Opt.* refers to whether the rule the subject is typed as using is the optimal one. *Per. Opt.* separates out those observations where *all* predictions of the participant are consistent with the optimal rule.

mistakes systematic, and how do they depend on the statistical structure of the data set?

4.3 What Types of Mistakes Do People Make?

In principle, participants can make two types of mistakes that result in suboptimal behavior. First, participants might completely ignore all observable variables. This represents a failure to utilize observable variables in the prediction task. We refer to this as a *non-conditioning error*. Second, participants might condition on the observable variables but do so in a suboptimal way. Depending on the statistical structure of the data set, this mistake might involve conditioning on an irrelevant variable or conditioning on only a subset of the relevant variables.[38] We refer to this as a *misalignment error*.

Furthermore, note that, from an ex-ante perspective, how the statistical features of the data set might influence the prevalence of these mistakes is not clear. For instance, noise can affect behavior in various ways. The absence of clear patterns in the data might increase the likelihood that participants erroneously identify nonexistent correlations. Alternatively, in an effort to simplify the environment, participants may ignore observable variables but make optimal predictions subject to this constraint. As we document in the next few sections, we surprisingly find limited evidence for either channel.

While the optimal rule describes the set of variables one should optimally condition on, we

[38] Conditioning on the correct set of observable variables but still doing so in a suboptimal way is also possible.

can identify the variables the participant actually conditions on from the prediction rule they are classified as using. For instance, the rule *G All* (for "Guess $S = 1$ for all light configurations") does not condition on any of the lights, whereas the rule *G w/ R* only conditions on the red light, and the rule *G w/ R or B* conditions on both lights. If the participant is typed as using *G All* when *G w/ R* is optimal, the participant will be classified as ignoring all relevant variables, and thus will be labeled as displaying a non-conditioning error. By contrast, if a participant is typed as using *G w/ R or B* when *G w/ R* is optimal, the participant will be classified as conditioning on an irrelevant variable, and thus will be labelled as displaying a misalignment error.

We present results using two figures. Figure 7 includes data sets where the optimal prediction rule is a function of only one of the observable variables. Figure 8 captures data sets where the optimal rule is a function of both observable variables. Non-conditioning errors describe participants ignoring all relevant variables in both figures. However, misalignment errors can take different forms based on the optimal number of variables to condition on. For the DAGs represented in Figure 7, misalignment errors describe participants conditioning on an irrelevant variable. For DAGs in Figure 8, misalignment errors describe participants conditioning on fewer variables than optimal, or conditioning on the right set of variables in an incorrect way. Finally, among the observations that are classified as corresponding to the optimal rule, the figures separate the subset of observations that match the optimal rule perfectly, denoted as Optimal (no errors) versus Optimal (w/ errors). We also develop a measure of "loss" associated with each category: for each observation, we compute its prediction accuracy and the decline it represents relative to the the optimal prediction rule. Note that because participants in the experiment were paid based on the accuracy of their guesses, the estimated loss for each category is directly linked to expected loss in payments (the latter is 25$ times the former). Each pie chart shows the breakdown into the categories discussed earlier and the average loss (L) within that category.

Before we describe the prevalence of different types of mistakes, we note that with respect to optimal behavior, both figures convey a consistent pattern. In all data sets, the majority (or close to it) of observations belong to one of two categories associated with optimal behavior. This finding shows participants are able to identify distinct patterns in a series of different data sets, which they then use to make predictions.

To describe mistakes, we start with Figure 7, which depicts DAGs (One Link and Chain), where conditioning on only one of the variables is optimal. First, in both One Link and Chain, we observe that both types of mistakes capture a non-negligible share of observations. However, in both cases, a non-conditioning error involving ignoring all relevant variables is the more damaging mistake,

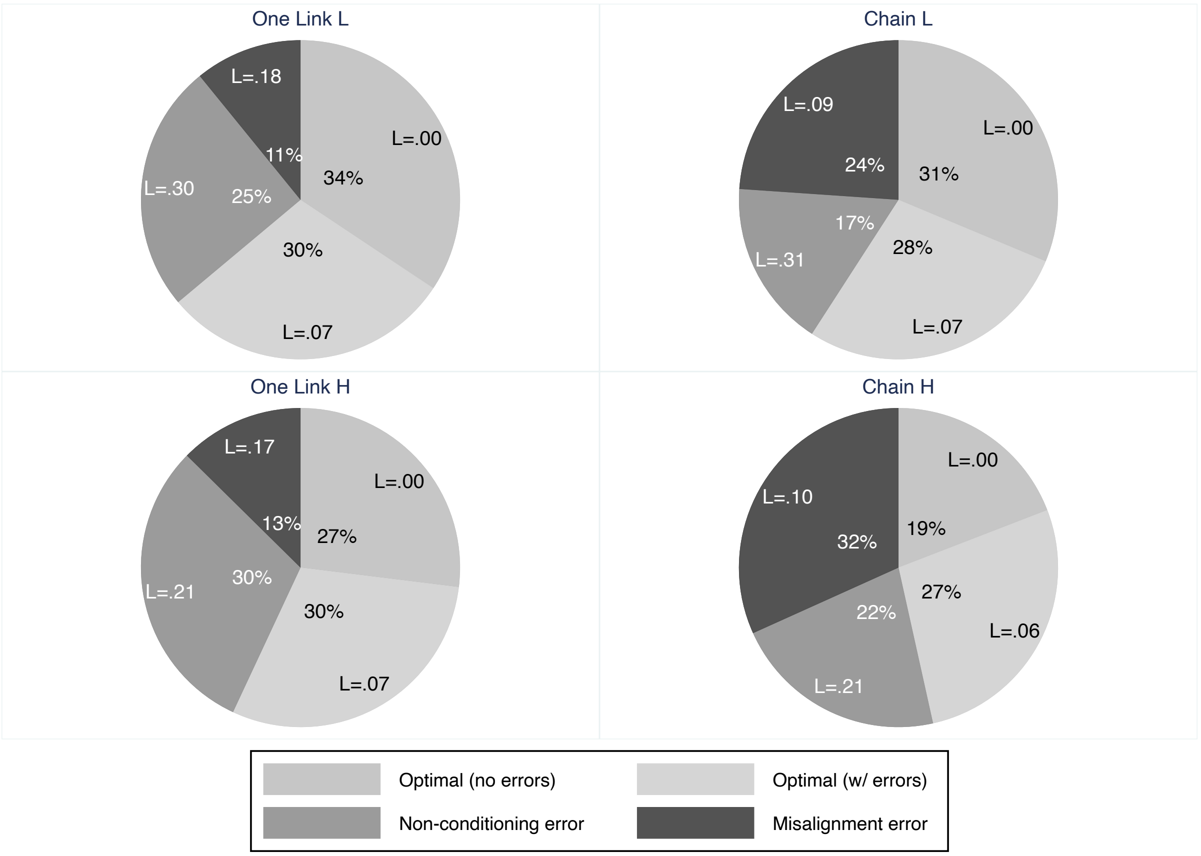


Figure 7: DAGs with one relevant variable: Mistakes and Associated Costs

Notes: See section 3 for detailed descriptions of each DAG. Pie charts in the top (bottom) row represent data from the low- (high-) noise environments and are labeled L (H). Percentages in each slice of the pie charts denote the relative share of that category. For each slice, the loss associated with such behavior, defined as the decline in guessing accuracy relative to optimal behavior, is also reported.

because it leads to a larger loss. Second, a common pattern emerges in One Link and Chain with respect to the effect of noise: both non-conditioning and misalignment errors become more likely in high-noise environments. Third, the contrast between One Link and Chain informs us about how the underlying correlation structure in the data set can impact the prevalence of different types of mistakes. Recall that the optimal prediction rule is G *w/* R in both One Link and Chain, but B and S are correlated only in the latter. The share conditioning on the irrelevant variable of B more than doubles in Chain relative to One Link, causing the increase associated with misalignment errors. This finding suggests a non-negligible share of participants struggle with understanding conditional independence and display a tendency to condition on any variable that correlates with the outcome, an issue we study further in the next section.

Next, we focus on Figure 8, which depicts DAGs (Common-consequence and Full), where conditioning on both variables is optimal. Here, misalignment errors are associated with conditioning on only one of the variables, or conditioning all variables but doing so in a suboptimal way. We first notice that both non-conditioning and misalignment errors are present in all cases. However, the share corresponding to non-conditioning errors is relatively larger in all cases. It captures at least 21 percent of observations (CC(AND)L and Full L) and at most 43 percent (CC(OR)H). Perhaps more importantly, the relative share of this mistake increases substantially when we move from low noise to high noise in all cases: by nine percentage points in CC(AND) and Full, and 20 percentage points in CC(OR).

In summary, we find evidence for both non-conditioning and misalignment errors. Thus, our results point toward heterogeneity in the types of mistakes people are prone to make. In future sections, we exploit our experimental design that allows us to observe predictions of the same participant across multiple data sets to study heterogeneity at the individual level more closely.

In addition, the results from this section inform us about how the environment (underlying structure of the data set) impacts the prevalence of these mistakes. We find (1) misalignment errors involving conditioning on an irrelevant variable are more likely when such variables are correlated with the outcome of interest; and (2) non-conditioning errors are more likely with high noise, particularly in settings where the optimal model involves making use of all relevant variables. However, as we show in the next section, non-conditioning errors are not associated with optimal behavior, subject to the constraint that predictions do not condition on the observable variables.

Result 3. *Both mistakes—non-conditioning and misalignment—are commonly observed. Conditioning on an irrelevant variable is more likely when the optimal prediction rule ignores variables correlated with the outcome. High noise increases the likelihood that relevant variables are missed.*

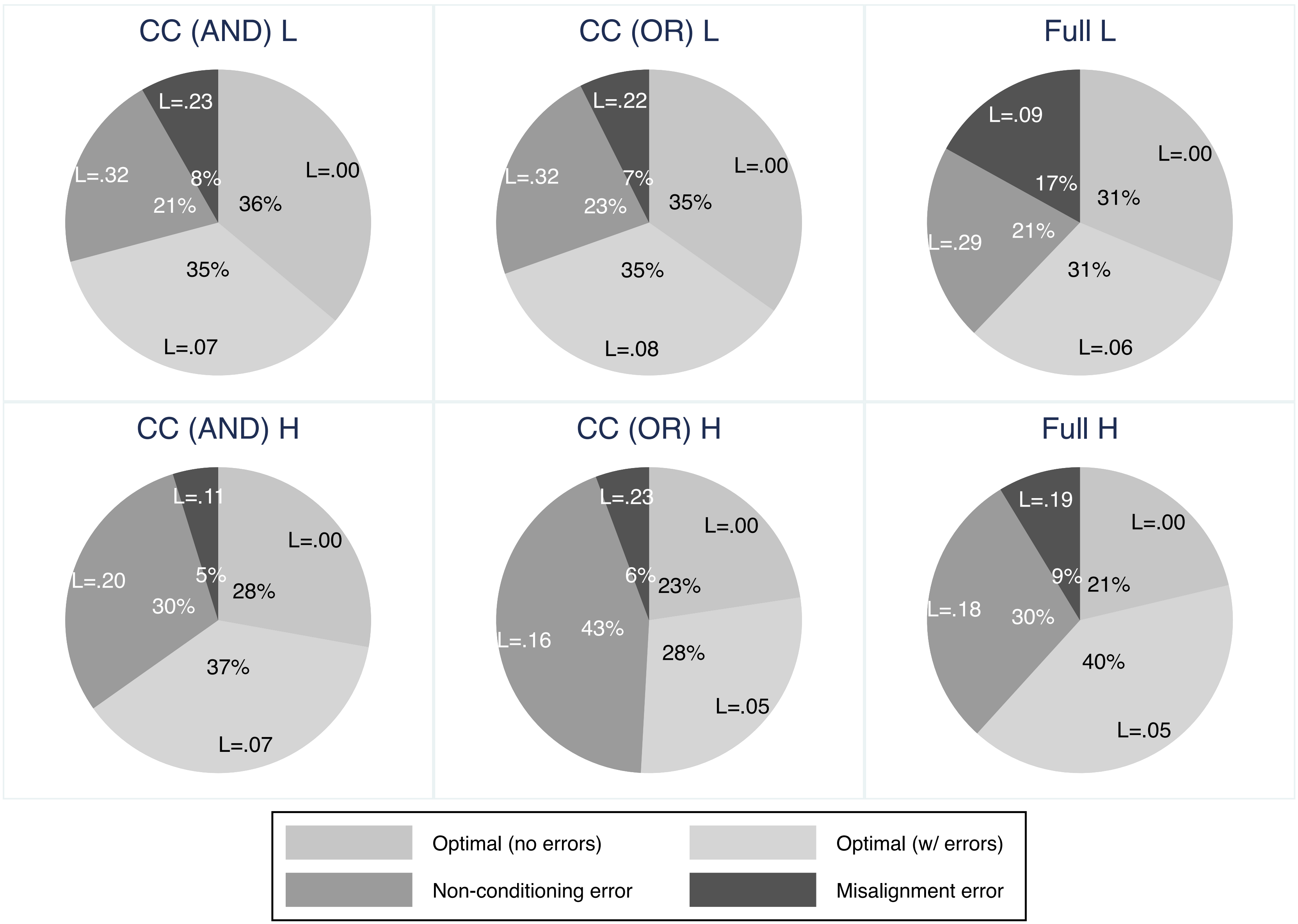


Figure 8: DAGs with two relevant variables: Mistakes and Associated Loss

Notes: See section 3 for detailed descriptions of each DAG. Pie charts in the top (bottom) row represent data from the low- (high-) noise environments and are labeled L (H). Percentages in each slice of the pie charts denote the relative share of that category. For each slice, the loss associated with such behavior, defined as the decline in guessing accuracy relative to optimal behavior, is also reported.

4.4 Why Do People Make Mistakes?

Our results indicate that most participants form mental models, as revealed by the prediction rules identified in the typing exercise. These rules, however, do not always align with the optimal ones. Can these deviations from optimal behavior be rationalized as arising from frictions or constraints in the learning process? For instance, what approaches to learning from data might lead to non-conditioning or misalignment errors? Understanding these connections could help us determine when certain mistakes are more likely to occur, which is the focus of our next analysis.

The way misalignment errors occur contrasts significantly with that of non-conditioning errors. In the One Link and Chain data sets, 81 percent of participants who display a misalignment error by conditioning on an irrelevant variable are classified as using the *G w/ R or B* or *G w/ R & B* prediction rules. These rules are constrained optimal in the sense that they perform best among the class of rules that condition on both lights. By contrast, only 11 percent of participants who display a non-conditioning error (i.e., ignore all relevant variables) are classified as using the rule *G All*, which corresponds to the optimal way to make predictions subject to the constraint that predictions do not condition on the lights. An overwhelming majority of such participants (87 percent), instead, are classified as displaying stochastic behavior.

Similar patterns are also observed in the Common Consequence and Full data sets: 96 percent of participants who display a misalignment error by conditioning on only one light are classified as using the *G w/ R* or *G w/ B* prediction rules, which perform best among the class of rules that condition on only one light. Only 16 percent of participants who display a non-conditioning error are classified as using the *G All* and, instead, 83 percent of such participants are classified as displaying stochastic behavior.

These results reveal the following pattern: participants who suffer from a misalignment error, that is, those who condition their predictions on the observable variables (the lights) but are misaligned in terms of how to do so, generally perform better than those who suffer from a non-conditioning error, that is, those who ignore all observable variables in the prediction task.[39]

Table 1 provides further insights on the qualitative difference between these two types of mistakes by contrasting optimality of predictions at different light configurations. In each

[39] The only exceptions to this pattern are CC (OR) H and Full H as seen in Figure 8, where the difference in loss between these two categories is low. Nonetheless, in every data set, the constrained optimal benchmark subject to a misalignment error achieves higher prediction accuracy than the constrained optimal benchmark subject to a non-conditioning error.

dataset, we isolate the light configuration $l := (R, B) \in \{(0,0),(1,0),(0,1),(1,1)\}$ that offers the "strongest evidence" for optimal behavior. We define this as the light configuration $l^\star$ at which $max\{p(S|l), 1 - p(S|l)\}p(l)$ takes the highest value. Note that, to identify $l^\star$, we combine information about how extreme (close to 0 or 1) $p(S|l)$ is, which represents the probability of the sound at each light configuration, with $p(l)$, the frequency with which this light configuration appears in the data set. Formally, $l^\star$ corresponds to the light configuration for which deviating from optimal behavior (by guessing randomly or in the opposite direction) would be most costly for the agent in terms of a decline in prediction accuracy.

Participants who are classified as using the optimal rule achieve a prediction accuracy of 95 percent for the $l^\star$ light configuration. Yet, this rate does not decline much for other light configurations. This finding is not surprising given our classification method, but is provided here nonetheless as a benchmark. Optimality rates corresponding to the different types of errors, reported in the second and third rows of Table 1, are more informative. Note that regardless of whether an agent suffers from a misalignment or non-conditioning error, it is still possible to achieve 100 percent optimality at $l^\star$.[40] Therefore, the table highlights the key differences between these two errors as observed in our data: (1) Those displaying misalignment errors achieve high optimality rates close to 90 percent for the light configuration for which evidence is strongest, but perform much worse with the other light configurations; (2) those displaying non-conditioning errors perform worse than other groups even with the light configuration for which evidence is strongest, but even in this group, optimality rates are higher with this light configuration relative to other light configurations.

Overall, these results suggest that when misalignment errors occur, they are relatively well calibrated in the sense that they generate predictions that are highly accurate in the cases—realizations of observables—that matter the most. Suboptimal predictions are much more likely to be observed in cases that are less common, or provide more mixed evidence on what constitutes optimal behavior. We note that this last observation also applies to non-conditioning errors even though substantial deviation from optimal behavior is observed in all cases for this category. Taken together, these results suggest the mental models participants form are responsive to the strength of evidence provided to them. In addition, misalignment and non-conditioning appear to be qualitatively very different types of mistakes in the sense that the former is closer to constrained optimal behavior, whereas the latter is often associated with stochastic behavior.

These results indicate that participants are more likely to make optimal predictions for $l^\star$, the light configuration with the strongest evidence. This finding supports the idea that even those

[40] Indeed, this is precisely how the constrained optimal benchmark subject to each error is achieved.

Table 1: Prediction Optimality

	One Link and Chain Data Sets		Common Consequence and Full Data Sets	
	Most evidence	Others	Most evidence	Others
Optimal	.95	.94	.95	.92
Misalignment error	.90	.69	.86	.64
Non-conditioning error	.67	.54	.69	.50

Notes: Misalignment refers to conditioning on variables in a suboptimal way. Non-conditioning error refers to ignoring variables altogether. "Most evidence" denotes the light configuration in each data set for which deviating from optimal behavior (by guessing randomly or in the opposite direction) is most costly; "Others" denotes all other light configurations.

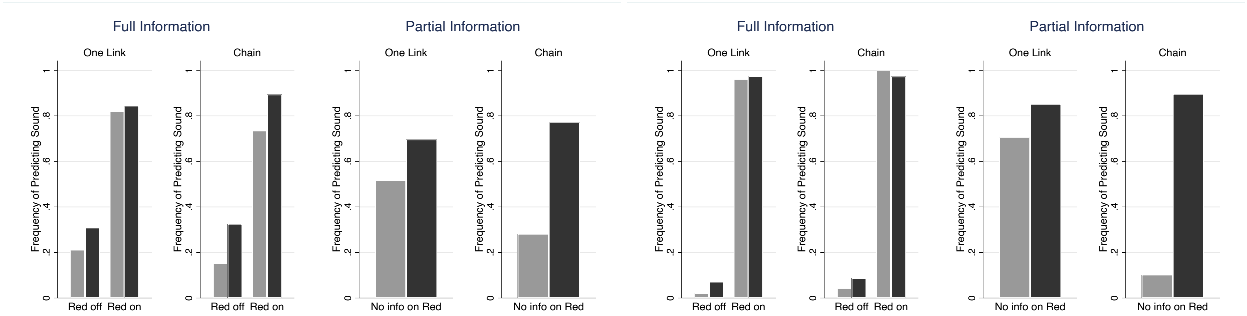
participants who cannot form optimal mental models are able to extract some information from the data sets to perform relatively well conditional on $l^\star$. From this perspective, participants who are classified as using the optimal rule distinguish themselves by achieving high prediction accuracy not just with $l^\star$, but across all possible light configurations l. This observation suggests they monitor the frequency of sound for all l. In fact, as we demonstrate in the next section, many of these participants form a much more sophisticated understanding of the statistical relationships between the variables that cannot be reduced to just learning about $p(S|l)$, the probability of the sound conditional on l.

Understanding of conditional independence

Consider the One Link and Chain DAGs. For a fixed level of noise, the values of $p(S|l)$ are the same in both DAGs. This translates into the same optimal behavior for each light configuration, as represented in the two left-most graphs of Figure 5, and the same optimal prediction rule, *G w/ R*. However, B is correlated with S in Chain but not in One Link. Now consider predictions for the additional tasks (rounds 28-36 of part 2) of the *Unspecified Prediction* treatment, where subjects were only provided with partial information on the status of the lights (R and B). Somebody who keeps track of only $p(S|l)$ (or the optimal prediction conditional on each l) would behave similarly in both One Link and Chain. However, optimal behavior for a prediction task in which only the status of B is known is different in these two DAGs. This requires an understanding of $p(S|B)$.

Figure 9 depicts the degree to which predictions on S condition on B in the presence of information on R (panels titled "Full Information") versus in the absence of information on R (panels "Partial Information").[41] Panel (a) of the figure reports the frequency of guessing $S = 1$, separated

[41] Online Appendix E presents more analysis on these additional rounds and shows the frequency of optimal pre-



(a) All Subjects

(b) Subjects Using Optimal Prediction Rule

Figure 9: Conditioning on the Blue Light in the Presence or Absence of Information on Red Light

Notes: Bar colors differentiate between when the blue light is on (black) vs. off (gray). In rounds 1-27, subjects are informed about the status of both lights. In rounds 28-36, subjects are informed about the status of only one light or no lights. In these rounds, we focus on rounds where only the status of the blue light is revealed. Panel (b) focuses on participants who are classified as using the optimal prediction rule in One Link and Chain in both high- and low-noise conditions. Graphs aggregate over low- and high-noise conditions.

by the status of B (and also R if observed) for all participants. Whereas behavior looks somewhat similar between One Link and Chain with full information (despite more evidence of conditioning on B in Chain), differences when the status of R is unobserved are large. For instance, the frequency of guessing $S = 1$ increases by 49 percentage points from $B = 0$ to $B = 1$ in Chain, while this increase is only 18 percent in One Link (p-value of < 0.001 and p-value of $= 0.001$, respectively). Panel (b) of the figure reproduces the same analysis, focusing on a subset of participants corresponding to 24 percent of the population who are classified (based on their behavior with full information) to be using the optimal prediction rule in all four data sets corresponding to the One Link and Chain DAGs.[42] For this subset of participants (who are by definition using the same optimal prediction rule with full information), differences in behavior in the later rounds are even more stark. The frequency of guessing $S = 1$ increases by 85 percent from $B = 0$ to $B = 1$ in Chain, while this increase is only 4 percent in One Link (p-value of < 0.001 and p-value of $= 0.701$, respectively).

These results suggest participants who are classified as optimal have mental models that do not solely rely on $p(S|l)$. Participants display sophistication, which suggests an understanding of conditional independence; namely, their behavior is consistent with an understanding that although B is correlated with S, the two variables are independent when conditioning on R. The mental model of these participants makes use of the correlation between B and S in the absence of any

diction declines with partial information. This finding suggests limitations on learning from data sets for at least some participants.

[42] Therefore, the left graph of panel (b) depicting behavior with full information is by construction very close to optimal and similar between Chain and One Link.

Table 2: Notes and Mistakes in One Link and Chain

	Share	Deterministic	Optimal (no errors)	Optimal (w/ errors)	Non-conditioning error (Ignores Relevant)	Misalignment error (Cond. on Irrelevant)
Codes All Data	.21	.73	.24	.34	.30	.12
Summarizes Frequency	.25	.95	.53	.35	.06	.06
Identifies Correlations	.26	.88	.28	.27	.13	.32
Other	.28	.62	.09	.21	.44	.26

Notes: "Codes All Data" consists of notes that would allow full replication of the data set (including order of the trials). "Summarizes Frequencies" consists of notes that summarize empirical frequencies of different kinds of trials (R, B, S). All other cases are classified under "Other", which is further separated based on whether the participant identifies correlations and causal relationships between different variables.

information on R, but also ignores this correlation in the presence of information on R.

Result 4.

(a) Most participants classified as using the optimal rule display understanding of conditional independence.

(b) Misalignment errors indicate constrained optimal behavior, because they yield high prediction accuracy for the most informative light configuration.

(c) Non-conditioning errors are associated with low prediction optimality across all light configurations and are strongly linked to stochastic behavior.

Different approaches to learning from data sets

The previous result suggests different ways in which participants learn from data sets and the mental models that they form as result. Our analysis so far has relied exclusively on participants' predictions, but in our experiment, participants take *notes* that they then use during the prediction task. Thus, these *notes-to-self* provide additional perspective on how participants examine and learn from data sets. We now study the types of notes participants take, how they relate to behavior in the prediction task, and discuss how the evidence arising from the notes is consistent with earlier results.[43] Here, we focus on the four data sets associated with the One Link and Chain DAGs, where both types of mistakes—non-conditioning and misalignment—are commonly observed.[44]

[43] Details on how the notes were coded are provided in Online Appendix F, which includes the protocol that two research assistants were provided to code the text in the notes.

[44] Online Appendix E provides additional analysis for the other data sets.

Table 2 separates observations (each unit being a participant in a data set) into four groups based on the style of notes associated with it. The first group corresponds to observations where participants' notes attempt to transcribe all the information in the data set (status of all variables in all trials in order). The second group corresponds to observations where participants' notes summarize information on the frequency of different kinds of trials. All observations in the remaining two groups, by definition, are associated with notes that include limited numerical information on the data set. In the third group, however, participants verbally describe correlations or causal relationships between different variables. The fourth group encompasses all other observations where no discernible numerical or verbal information about statistical patterns in the data set is highlighted.[45] These four groups represent roughly similar shares of observations ranging from 21 to 28 percent.

Table 2 reveals the following insights. First, using a deterministic prediction rule is associated with specific styles of note-taking, in particular, reporting key frequencies or directly identifying correlations in the data. This finding suggests that some participants study data sets directly with the goal of identifying patterns they will later use in the prediction task. Failing to do this at the note-taking stage makes it more difficult for participants to use such patterns at the prediction stage, even if notes contain sufficient information to reveal these patterns (e.g., in those cases where notes code all the data).[46]

These patterns also suggest distinct mechanisms through which stochastic behavior might arise in our experiment, particularly taking into account that stochastic behavior is associated with ignoring relevant variables. Some participants who code all the data might sample from these observations to make predictions. For example, when asked whether the sound is on when both lights are on, they may be finding a similar trial in their notes (where both lights were on) and base their prediction on whether the sound was on in that specific trial. Alternatively, stochastic behavior could be a response to lack of information, as with the last category of notes, and could also result from difficulties interpreting the complex information in the first category of notes. It is worth noting that we find limited evidence for "classical" probability matching behavior: Only five percent of participants in the second category—who specifically code the probability of the sound—display stochastic behavior.

[45] Examples of classification of notes are provided in Online Appendix E.

[46] Overall, coding all the data appears to be a highly ineffective method of note-taking. It is costly in terms of time spent taking notes and making predictions. In general, spending more time on the prediction tasks is associated with higher accuracy and optimality, but the same is not true for time spent taking notes. See Tables 13 and 14 in Online Appendix E.

Second, note-taking in the form of summarizing frequencies is by far associated with the highest levels of optimal behavior: 53 percent of these observations correspond perfectly to optimal behavior, and a remaining 35 percent are associated with noisy optimal behavior. Third, different note-taking styles are associated with different types of mistakes. When notes code all the data or—at the opposite end—carry little information as in the last category, the likelihood of non-conditioning errors, that is, missing relevant variables, is high (30 or 44 percent, respectively). By contrast, when notes verbally identify correlations in the data, the likelihood of misalignment errors that take the form of conditioning on an irrelevant variable in these data sets is high (32 percent).

Taken together, there results display heterogeneity in how participants learn from data sets and are consistent with some of the patterns described in Result 4. Some note-taking styles clearly suggest participants are deliberately looking for patterns in the data. They do so by either summarizing key statistical properties or by qualitatively identifying and verbally reporting correlations among variables. Others are unable or choose not to engage with the data in the same way, which proves costly at the prediction stage.

Can mistakes be rationalized with subjective priors?

The evidence we have used to understand the type of mental models that participants form relies exclusively on the information (i.e., data sets) that participants were provided. For participants whose behavior is consistent with the optimal rule, this approach seems to be a reasonable approximation. But for participants who display mistakes, such deviations may be explained by participants having strong prior beliefs about what constitutes optimal behavior. To clarify, consider a Bayesian agent who begins with a prior belief that the likelihood of $S = 1$ is nearly 1, irrespective of the light configuration. A data set consisting of 27 observations might not provide sufficiently strong evidence to dissuade the agent from predicting $S = 1$ for any light configuration (which requires the posterior likelihood of $S = 1$ to fall below 0.5). Such an agent would be classified as ignoring all relevant variables, that is, displaying a non-conditioning error. Starting with other priors could also result in misalignment errors, behavior that conditions on irrelevant variables or conditions on fewer variables than is optimal. Although the abstract framing of the experimental design was intended to minimize such "home grown" priors, we cannot conceptually rule out that participants nonetheless bring such priors to the experiment. We can, however, study the extent to which the observed behavior of any participant can be rationalized with *some* prior (within a reasonable class). One way of analyzing the extent to which this is possible is to compute the minimal number of predictions a participant would need to change such that their behavior can be fully rationalized by a prior.[47] The key idea that this analysis builds on is that regardless of an agent's prior, the optimal response to information should be monotonic in the strength of the evidence provided in the data set for a large set of priors.[48]

This analysis is presented in Online Appendix E. Here, we summarize the main takeaways. Although substantial heterogeneity exists, participants who are classified as displaying misalignment errors are closer to Bayesian behavior (with *some* prior) than participants who are classified as displaying non-conditioning errors. That is, the two mistakes manifest themselves in qualitatively very different ways in our data. Supporting findings from the previous section, non-conditioning errors are often associated with persistently stochastic behavior (mixing between predicting $S = 1$ and $S = 0$), which cannot be rationalized as optimal behavior for any prior.

4.5 Is Behavior at the Individual-Level Consistent across Data Sets?

Table 3: Behavior in One Link and Chain

		High Noise			
		Optimal	Optimal	Non-conditioning error	Misalignment error
		(no errors)	(w/ errors)	(Ignores relevant var.)	(Cond. on irrelevant var.)
	Optimal (no errors)	**.55**	.26	.07	.11
Low	Optimal (w/ errors)	.10	**.48**	.20	.21
Noise	Non-conditioning error	.03	.15	**.67**	.14
	Misalignment error	.09	.16	.21	**.54**

Notes: The table reports the likelihood of different categories of behavior in the high-noise data set of each DAG as a function of the category of behavior in the low-noise data set of the same DAG. For example, 67 percent of subjects who ignored a relevant variable in One Link L or Chain L also ignore a relevant variable in One Link H or Chain H.

Our previous analysis revealed several patterns in the mental models that participants form after examining data sets. However, this analysis primarily relied on observations at the *participant-data set* level. The interpretation of these findings varies based on whether the patterns are consistent

[47] This analysis is similar in spirit to Afriat (1973) critical cost efficiency index (CCEI), which measures the minimal degree to which budget constraints need to be adjusted to remove all GARP violations from a participant's collection of choices (thereby rendering their behavior rationalizable).

[48] Fixing the configuration of the lights, we posit that if it is optimal for an agent to predict $S = 1$ when the data set includes 10 rounds of this light configuration where $S = 1$ in 70 percent of cases, it should also be optimal for the agent to predict $S = 1$ when the data set includes 20 rounds of this light configuration where $S = 1$ in 80 percent of cases. We show that this is true for any prior characterized by the beta distribution.

for individuals across different data sets. A lack of consistency would suggest the information participants extract and how they use it depends on the structure of the data. On the other hand, evidence of consistency would indicate the heterogeneities identified at the *participant-data set* level are actually capturing distinct participant types. Overall, results reported in this section reveal a high level of consistency in how participants learn (or fail to learn) from data sets.

As a starting point, we decompose the variance in prediction optimality to variation across data sets and across participants: whereas the former accounts for 4 percent, the latter accounts for 96 percent! This finding indicates a high degree of correlation across data sets in a participant's ability to make optimal predictions.[49]

Furthermore, participants typically maintain the same note-taking approach across various data sets. For example, notes from 77 percent of participants are classified as being within the same group in *all* four data sets reported in Table 2.[50]

Consistency of behavior can also be observed with respect to other measures. Across the 10 data sets mentioned, (i) 41 percent of participants are *always* classified as using a deterministic prediction rule, whereas 16 percent are classified as using one less than half the time; and (ii) 34 percent of participants *always* condition on at least one of the observable variables, whereas 19 percent do so less than half the time.

More importantly, Table 3 shows that participants are likely to repeat the same mistakes across multiple data sets. The table focuses on the One Link and Chain data sets, where both types of mistakes—non-conditioning and misalignment—are commonly observed.[51] The table reports the likelihood of different categories of behavior in the high-noise data set of each DAG as a function of category of behavior in the low-noise data set of the same DAG. The diagonal values, highlighted in bold, correspond to full consistency. For instance, we see participants whose predictions perfectly match the optimal rule in the low-noise data set achieve the same outcome with a 55 percent likelihood in the high-noise data set. Alternatively, participants who display a non-conditioning (misalignment) error in the low-noise data set repeat the same mistake with a 67 (54) percent likelihood in the high-noise data set.

[49] Another way to see consistency is to look at whether a participant's prediction optimality exceeds the median value for that data set: across ten data sets, 15 percent of participants *always* achieve this, whereas 17 percent of participants *never* achieve it.

[50] Sixty-four percent of participants are classified as being within the same group in *all* ten data sets. The observed variation is mostly among the last two categories, which are more difficult to separate.

[51] Corresponding values for the Common Consequence and Full data sets in Table 12 of Online Appendix E reveal similar results.

Finally, we find that mistakes in One Link and Chain are predictive of mistakes in the other DAGs. For instance, participants who never exhibit a non-conditioning error in the four data sets associated with One Link and Chain have only a 10 percent likelihood of suffering from this mistake in the reminder six data sets where conditioning on both R and B is optimal. By contrast, those who have displayed a non-conditioning error at least once in data sets associated with One Link and Chain have a 53 percent likelihood of doing so in the remaining ones.

Result 5. *Across participants, we find heterogeneities in behavior, as captured by patterns identified in previous results. For a fixed participant, behavior tends to be consistent across data sets. Specifically, behavior in one data set is highly predictive of their behavior in another data set.*

Limited impact of learning

We observe limited learning effects across different data sets. Table 15 in Online Appendix E indicates only a minor decline in prediction accuracy and optimality in later data sets than in earlier ones. Furthermore, Table 16 in the same appendix suggests the likelihood of conditioning on a variable may be influenced by certain features of the previous data set. For instance, we find that participants are more likely to condition on a light if doing so in the previous dataset was optimal, although this effect, while statistically significant, is limited in scope.

Implications of consistent heterogeneity: Disagreement

Our results reveal substantial heterogeneity in how participants learn from data sets and their propensity to make different mistakes. So far, we have only highlighted the payoff consequences of deviating from optimal behavior directly for the participants, as reported in Figures 7 and 8 in terms of a decline in prediction accuracy. Such deviations can also have social consequences. Our results show how different participants, observing the same information, make conflicting inferences about what constitutes an optimal action. Such disagreements can be quantified. We define the disagreement rate among any pair of participants at the data set level as the likelihood that the participants disagree—make opposing predictions—based on the same information. Note that the disagreement rate is necessarily zero among any pair of participants who perfectly follow the optimal prediction rule. However, heterogeneity in how participants learn from data sets generates disagreements.

We summarize our analysis on disagreement rates here and report detailed results in Online Appendix E. First, at the aggregate level, we find evidence of significant disagreement for *all*

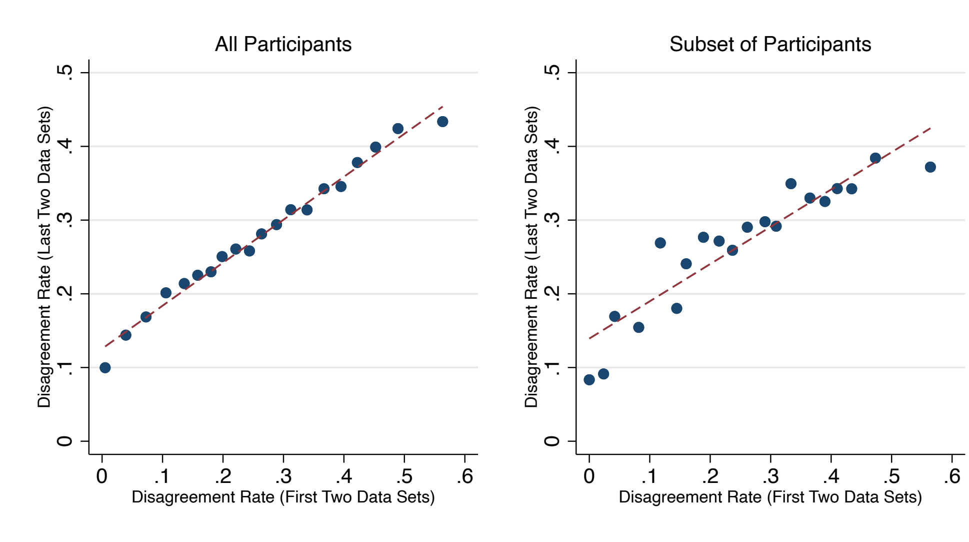


Figure 10: Disagreement Rates in Early and Late Data Sets

Notes: Figure shows binned scatter plots. Disagreement Rate is the frequency with which a pair of participants make opposing predictions about S after observing the same information on (R, B). The left panel includes all participants; the right panel focuses on a subset of participants (60 percent of the overall population) whose predictions on S are aligned in the absence of any information on (R, B).

light configurations and data sets where we observe at least three predictions for each participant. Second, we observe that disagreement rates increase with noise in each DAG.

Third, demonstrating the implications of observed heterogeneity across participants, we find disagreements among pairs of participants are largely predictable. Figure 14 depicts the average disagreement rate among pairs in the last two data sets as a function of disagreement rates in the first two data sets to display strong correlation between these two measures.[52] Finally, we find that our results on disagreement hold even when we focus on pairs of participants whose predictions align in the absence of information about the lights, suggesting that disagreement rates can be reliably interpreted as measures of polarization. By polarization, we mean divergent responses to new information, where participants who initially agree react differently when provided with information about the status of the lights.

Result 6. *Consistency of behavior across different data sets makes predicting who will disagree with whom possible.*

[52] To rule out that all disagreements are driven by stochastic predictions, Figure 20 in Online Appendix E reproduces these graphs, focusing only on participants who are classified as using a deterministic prediction rule.

5 Discussion

We experimentally study the types of mental models people adopt after examining data; namely, we study the inferences they make about the statistical relationships among variables in a set of observations. In doing so, we establish a variety of new findings regarding the types of models people form by observing data, how these models vary with the underlying statistical relationships, their optimality, and the structure of the notes participants leave for themselves.

One feature of our results is that for a given participant, the way they learn across data sets is highly consistent. The largest share of our participants display optimal behavior: 31% identify the correct corelation structure in 9 of the 10 data sets. These participants are able to extract sophisticated patterns from the data. In fact, 61% of these participant also exhibit an understanding of conditional independence. Specifically, these participants make optimal decisions about whether to condition their predictions on a variable, based on the availability of information regarding confounding variables, effectively ignoring irrelevant correlations when necessary.

While optimal behavior is remarkably consistent across data sets for a given individual, the prevalence of different types of mistakes—though still predictable at the individual level—also depend on features of the data set.

Non-conditioning errors, failures to make use of *any* of the existing correlations in the data, are observed 26% of the time across all data sets. This is the most common error across all data sets except *Chain*, where it is still observed 19% of the time. Weakening the correlation between the variables increases this mistake on average by 10 percentage points. Interestingly, many participants who suffer from this mistake take detailed time-consuming notes on the data sets they observe, yet they make predictions that do not reflect the information they collected. Furthermore, these errors are associated with stochastic behavior 85% of the time; thus, such behavior is not optimal within the class of models that do not condition on observable variables. Overall, our findings suggest that these participants either struggle to identify statistical patterns in the data or fail to utilize them optimally, even when they successfully record such information.

We also observe *misalignment errors.* This mistake, observed 13% of the time, involves conditioning on observable variables but failing to do so optimally. Importantly, this mistake occurs most often in the presence of confounding variables, rising to 28% in such cases. The evidence indicates that participants who tend to make this mistake have a good understanding of the most consequential statistical patterns in data sets. Specifically, we find that the predictions of these participants are highly accurate in the contingencies that matter the most—configuration of observables where deviation from optimal behavior would be most costly. In this sense, participants who make misalignment errors are engaging in a form of constrained optimal behavior, which stands in sharp contrast to those who display non-conditioning errors. The fact that these errors are most often observed among participants who only keep coarse qualitative summaries of the data supports this view.

A recent literature is interested in persuasion via narratives (see, e.g., Schwartzstein, Sunderam (2021, 2022); Aina (2023)). Since a substantial fraction of subjects do not identify the correct correlation structure on their own, our results suggest scope for such persuasion. For instance, subjects who suffer from *nonconditioning* or *misalignment errors* errors could be presented with models that explain the data better; including some that are not the optimal model. That is, even those who struggle to grasp the true correlation structure on their own may still be capable of recognizing and adopting alternative models that align more closely with the data. Exactly who can be persuaded to change their model and under what conditions is an interesting question for future work.

More broadly, a key implication of the systematic heterogeneity in models adopted by participants is that, even when presented with identical observations, people will often disagree on what is optimal behavior. In our data, the probability that two participants disagree in their predictions in any given data set is 27%. Moreover, the disagreement rate among a pair of participants in one data set is predictive of their disagreement rate in another data set.[53] Studying disagreement arising from different interpretations of the same information can be important for shedding light on diverse phenomena such as ideological polarization or high trading volumes in financial markets.[54]

References

Afriat, S. N. (1973), 'On a system of inequalities in demand analysis: an extension of the classical method', *International economic review* pp. 460–472.

Aina, C. (2023), 'Tailored stories', *Working Paper* .

[53] For example, with a fixed correlation structure, the disagreement rate between a pair of participants across low-noise and high-noise data sets shows a correlation of 0.53.

[54] We document, for example, directional disagreement on how to bet on certain events among subjects who were given identical information. This connects to the large literature in finance, going back to Miller (1977) and Harrison & Kreps (1978) studying sustained disagreement between optimists ("bulls") and pessimists ("bears") in financial markets.

Akerlof, G. A. & Shiller, R. J. (2010), *Animal spirits: How human psychology drives the economy, and why it matters for global capitalism*, Princeton university press.

Ali, S. N., Mihm, M., Siga, L. & Tergiman, C. (2021), 'Adverse and advantageous selection in the laboratory', *American Economic Review* **111**(7), 2152–78.

Ambuehl, S. & Thysen, H. C. (2024), 'Choosing between causal interpretations: An experimental study', *Working Paper* .

Andre, P., Haaland, I., Roth, C. & Wohlfart, J. (2021), 'Narratives about the macroeconomy', *Working Paper* .

Aoyagi, M., Fréchette, G. R. & Yuksel, S. (2024), 'Beliefs in repeated games: An experiment', *American Economic Review* .

Araujo, F. A., Wang, S. W. & Wilson, A. J. (2021), *American Economic Journal: Microeconomics* **13**(4), 1–22.

Barron, K. & Fries, T. (2024), 'Narrative persuasion', *Working Paper* .

Bohren, J. A. & Hauser, D. N. (2021), 'Learning with heterogeneous misspecified models: Characterization and robustness', *Econometrica* **89**(6), 3025–3077.

Bramley, N. R., Gerstenberg, T., Mayrhofer, R. & Lagnado, D. A. (2018), 'Time in causal structure learning.', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **44**(12), 1880.

Bursztyn, L., Rao, A., Roth, C. & Yanagizawa-Drott, D. (2023), 'Opinions as facts', *The Review of Economic Studies* **90**(4), 1832–1864.

Cason, T. N. & Plott, C. R. (2014), 'Misconceptions and game form recognition: Challenges to theories of revealed preference and framing', *Journal of Political Economy* **122**(6), 1235–1270.

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

Charness, G. & Levin, D. (2009), 'The origin of the winner's curse: a laboratory study', *American Economic Journal: Microeconomics* **1**(1), 207–236.

Dal Bó, E., Dal Bó, P. & Eyster, E. (2018), 'The demand for bad policy when voters underappreciate equilibrium effects', *The Review of Economic Studies* **85**(2), 964–998.

Eliaz, K. & Spiegler, R. (2020), 'A model of competing narratives', *American Economic Review* **110**(12), 3786–3816.

Enke, B. (2020), ‘What you see is all there is’, *The Quarterly Journal of Economics* **135**(3), 1363–1398.

Enke, B. & Zimmermann, F. (2019), ‘Correlation neglect in belief formation’, *The Review of Economic Studies* **86**(1), 313–332.

Esponda, I. & Pouzo, D. (2016), ‘Berk–nash equilibrium: A framework for modeling agents with misspecified models’, *Econometrica* **84**(3), 1093–1130.

Esponda, I. & Vespa, E. (2014), ‘Hypothetical thinking and information extraction in the laboratory’, *American Economic Journal: Microeconomics* **6**(4), 180–202.

Esponda, I. & Vespa, E. (2018), ‘Endogenous sample selection: A laboratory study’, *Quantitative Economics* **9**(1), 183–216.

Esponda, I. & Vespa, E. (2021), ‘Contingent thinking and the sure-thing principle: Revisiting classic anomalies in the laboratory’, *Working Paper* .

Esponda, I., Vespa, E. & Yuksel, S. (2024), ‘Mental models and learning: The case of base-rate neglect’, *American Economic Review* **114**(3), 752–782.

Eyster, E. & Weizsäcker, G. (2010), ‘Correlation neglect in financial decision-making’, *Working Paper* .

Fernbach, P. M., Darlow, A. & Sloman, S. A. (2011), ‘Asymmetries in predictive and diagnostic reasoning.’, *Journal of Experimental Psychology: General* **140**(2), 168.

Fudenberg, D., Romanyuk, G. & Strack, P. (2017), ‘Active learning with a misspecified prior’, *Theoretical Economics* **12**(3), 1155–1189.

Fudenberg, D. & Vespa, E. (2019), ‘Learning theory and heterogeneous play in a signaling-game experiment’, *American Economic Journal: Microeconomics* **11**(4), 186–215.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T. & Danks, D. (2004), ‘A theory of causal learning in children: causal maps and bayes nets.’, *Psychological review* **111**(1), 3.

Hanna, R., Mullainathan, S. & Schwartzstein, J. (2014), ‘Learning through noticing: Theory and evidence from a field experiment’, *The Quarterly Journal of Economics* **129**(3), 1311–1353.

Harrison, J. M. & Kreps, D. M. (1978), ‘Speculative investor behavior in a stock market with heterogeneous expectations’, *The Quarterly Journal of Economics* **92**(2), 323–336.

Heidhues, P., Kőszegi, B. & Strack, P. (2018), ‘Unrealistic expectations and misguided learning’, *Econometrica* **86**(4), 1159–1214.

Kendall, C. W. & Oprea, R. (2024), ‘On the complexity of forming mental models’, *Quantitative Economics* **15**, 175–211.

Le Pelley, M. E., Griffiths, O. & Beesley, T. (2017), ‘Associative accounts of causal cognition’, *The Oxford handbook of causal reasoning* pp. 13–28.

Martin, D. & Muñoz-Rodriguez, E. (2019), ‘Misperceiving mechanisms: Imperfect perception and the failure to recognize dominant strategies’, *Working Paper* .

Martínez-Marquina, A., Niederle, M. & Vespa, E. (2019), ‘Failures in contingent reasoning: The role of uncertainty’, *American Economic Review* **109**(10), 3437–74.

Miller, E. M. (1977), ‘Risk, uncertainty, and divergence of opinion’, *The Journal of finance* **32**(4), 1151–1168.

Ngangoué, M. K. & Weizsäcker, G. (2021), ‘Learning from unrealized versus realized prices’, *American Economic Journal: Microeconomics* **13**(2), 174–201.

Pearl, J. (2009), *Causality*, Cambridge university press.

Rottman, B. M. (2017), ‘The acquisition and use of causal structure knowledge’, *The Oxford handbook of causal reasoning* pp. 85–114.

Rottman, B. M. & Hastie, R. (2014), ‘Reasoning about causal relationships: Inferences on causal networks.’, *Psychological bulletin* **140**(1), 109.

Schwartzstein, J. & Sunderam, A. (2021), ‘Using models to persuade’, *American Economic Review* **111**(1), 276–323.

Schwartzstein, J. & Sunderam, A. (2022), ‘Shared models in networks, organizations, and groups’, *Working Paper* .

Spiegler, R. (2020), ‘Behavioral implications of causal misperceptions’, *Annual Review of Economics* **12**, 81–106.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J. & Blum, B. (2003), ‘Inferring causal networks from observations and interventions’, *Cognitive science* **27**(3), 453–489.

Weisberg, D. S., Choi, E. & Sobel, D. M. (2020), ‘Of blickets, butterflies, and baby dinosaurs: childrens diagnostic reasoning across domains’, *Frontiers in Psychology* **11**, 2210.

Online Appendix for

Extracting Models From Data Sets: An Experiment Using Notes-to-Self

Guillaume Fréchette Emanuel Vespa Sevgi Yuksel

CONTENTS:

A Details of Each Data Sets

In this section, we describe the procedure with which we determined the 11 data sets presented to the subjects.

First, we picked the 5 DAGs (Directed Ayclic Graphs) described in Figure 3 (with two versions of Common Consequence corresponding to AND and OR conditions) as described in Section 3.2. These cover all possible DAGs with three variables (R, B, and S) where one variable (S) is fixed not to cause any of the other variables.[55] For all DAGs where there is some causal relationship between the variables, we also varied the strength of these causal connections as described in Section 3.2.[56]

For each case (Low or High noise) of each DAG (separating AND and OR conditions for Common Consequence), the following parameters remain free:

No Correlation. Three free parameters: probability of $B = 1$, $R = 1$ and $S = 1$.

One Link (Low and High noise). Two free parameters: probability of $B = 1$ and $R = 1$.

Chain (Low and High noise). One free parameter: probability of $B = 1$.

CC AND (Low and High noise). Two free parameters: probability of $B = 1$ and $R = 1$.

CC OR (Low and High noise). Two free parameters: probability of $B = 1$ and $R = 1$.

Full (Low and High noise). One free parameter: probability of $B = 1$.

These free parameters were picked to achieve two goals:

1. Keep probability $S = 1$ is on around 62 percent.
2. Increase identification across the 11 different cases: Namely, increase cost of using the optimal prediction rule from one case in another while preserving some variation in the data set.

Table 4 lists the parameters chosen for each case. The table also reports the implied probability of each light and sound configuration (R, B, S) given these parameters. Finally, the table also includes the finite sample approximation (over 27 trials) that was shown to the subjects.

[55] In the actual implementation of the experiment, we also vary which lights are labeled as B versus B allowing us to create further variation in the DAGs presented to subjects.

[56] For instance, in the One Link Low Noise condition $p(S = 1 \mid R = 1) = p(S = 0 \mid R = 0) = 0.90$. This value was changed to 0.80 in the high noise condition.

Two characteristics of the parameterization as reflected in the finite sample approximation are worth bringing to attention. (1) In the No Correlation data set, there are no trials where both $R = 1$ and $B = 1$. This is due to the fact that the parameters were chosen to increase likelihood of observing both $R = 0$ and $B = 0$, a configuration in which optimal prediction on S for this case is uniquely different from all others, helping with identification. (2) In the Common Consequence (OR) High Noise data set, there are no trials where both $R = 0$ and $B = 0$. This is due to the fact that the probability of $R = 1$ and $B = 1$ needed to be high enough to ensure probability of the sound remained around 62 percent with high noise.

Table 4: Characteristics of Each Data Set

	Probability of each event			Probability for each configuration of (B, R, S)							
	p(B = 1)	p(R = 1)	p(S = 1)	(0, 0, 0)	(0, 0, 1)	(1, 0, 0)	(1, 0, 1)	(1, 1, 0)	(1, 1, 1)	(0, 1, 0)	(0, 1, 1)
No Correlation	0.10	0.10	0.62	0.31	0.50	0.03	0.06	0.00	0.01	0.03	0.06
	3	2	17	8	14	1	2	0	0	1	1
One Link, Low Noise	0.50	0.65	*0.62*	0.16	0.02	0.16	0.02	0.03	0.29	0.03	0.29
	14	18	17	4	0	4	1	1	8	1	8
One Link, High Noise	0.50	0.70	*0.62*	0.12	0.03	0.12	0.03	0.07	0.28	0.07	0.28
	14	19	17	3	1	3	1	2	8	2	7
Chain, Low Noise	0.70	*0.66*	*0.63*	0.24	0.03	0.06	0.01	0.06	0.57	0.00	0.03
	19	18	17	6	1	2	0	2	15	0	1
Chain, High Noise	0.80	*0.68*	*0.61*	0.13	0.03	0.13	0.03	0.13	0.51	0.01	0.03
	22	19	17	3	1	3	1	4	14	0	1
CC (AND), Low Noise	0.75	0.85	*0.61*	0.03	0.00	0.10	0.01	0.06	0.57	0.19	0.02
	20	23	16	1	0	3	0	2	15	5	1
CC (AND), High Noise	0.80	0.85	*0.61*	0.02	0.01	0.10	0.02	0.14	0.54	0.14	0.03
	22	24	17	0	0	2	1	4	15	4	1
CC (OR), Low Noise	0.54	0.20	*0.62*	0.33	0.08	0.03	0.31	0.00	0.11	0.01	0.13
	12	7	17	9	2	1	8	0	3	0	4
CC (OR), High Noise	0.50	0.25	*0.62*	0.31	0.17	0.05	0.27	0.00	0.08	0.02	0.10
	10	6	17	8	5	1	7	0	2	1	3
Full, Low Noise	0.50	*0.50*	*0.62*	0.36	0.09	0.00	0.05	0.00	0.45	0.00	0.05
	13	13	17	10	3	0	1	0	12	0	1
Full, High Noise	0.35	*0.41*	*0.62*	0.33	0.19	0.01	0.06	0.01	0.27	0.02	0.11
	9	11	17	9	5	0	2	0	7	1	3

Notes: The first three columns (labelled as "Probability of each event") describes the probability with which the blue and red lights as well a the sound is on. If the probability is reported in italics, it is not a free parameter but is derived from other primitives given the DAG structure as described in Section 3.2 and earlier in this Online Appendix. The last 8 columns denote the probability with which the different light and sound configurations occur. Integers underneath each value show the finite sample approximation (27 trials in total) shown to the subjects.

B Description of Prediction Rules

Table 5: Prediction Rules

	Guess By Light Configuration				Conditioning on Lights	
Rule	$R = 1$, $B = 1$	$R = 1$, $B = 0$	$R = 0$, $B = 1$	$R = 10$, $B = 0$	Red?	Blue?
G All	1	1	1	1	0	0
G w/ R	1	1	0	0	1	0
G w/ B	1	0	1	0	0	1
G w/ R & B	1	0	0	0	1	1
G w/ R or B	1	1	1	0	1	1
G Never	0	0	0	0	0	0
G w/ not R	0	0	1	1	1	0
G w/ not B	0	1	0	1	0	1
G w/ not R or not B	0	1	1	1	1	1
G w/ not R & not B	0	0	0	1	1	1
G 0100	0	1	0	0	1	1
G 0010	0	0	1	0	1	1
G 1101	1	1	0	1	1	1
G 1011	1	0	1	1	1	1
G 1001	1	0	0	1	1	1
G 0110	0	1	1	0	1	1

Notes: These 16 rules represent all possible deterministic rules in this setting. A rule is referred to as conditioning on a light, if conditioning on the status of the other light, guesses do change with the status of this light. R and B denote the status of the red and blue lights. To facilitate reading of the table, consider the the following examples: The first rule, abbreviated as *G All*, guesses the sound to be on for all light configurations, and thus does not condition on either the red or the blue light. The fourth rule, abbreviated as *G w/ R & B*, guesses the sound to be on only when both red and blue lights are on. When blue is off, the guess independent of the status of the red light, but when blue is on, the guess depends on the status of the red light. A similar argument shows that this rule conditions on *both* lights.

C Further Analysis on the Aggregate Level

Table 6: Prediction Accuracy by Data Set and Deterministic Prediction Rule (%)

Parametrization	Prediction Rule			
	G All	*G w/ R*	*G w/ R & B*	*G w/ R or B*
No Correlation	**62**	40	38	43
One Link High	62	**80**	59	71
One Link Low	62	**90**	64	76
Chain High	61	**80**	78	70
Chain Low	63	**90**	88	84
CC(AND) High	61	70	**80**	63
CC(AND) Low	61	73	**90**	64
CC(OR) High	62	54	45	**75**
CC(OR) Low	62	60	49	**88**
Full High	61	72	63	**77**
Full Low	61	86	82	**90**

Notes: *G All*, guesses that the machine makes a sound for all light configurations. *G w/ R*, guesses the sound only when the red light is on. *G w/ R & B* guesses the sound when both lights are on. *G w/ R or B* guesses the sound when either light is on.

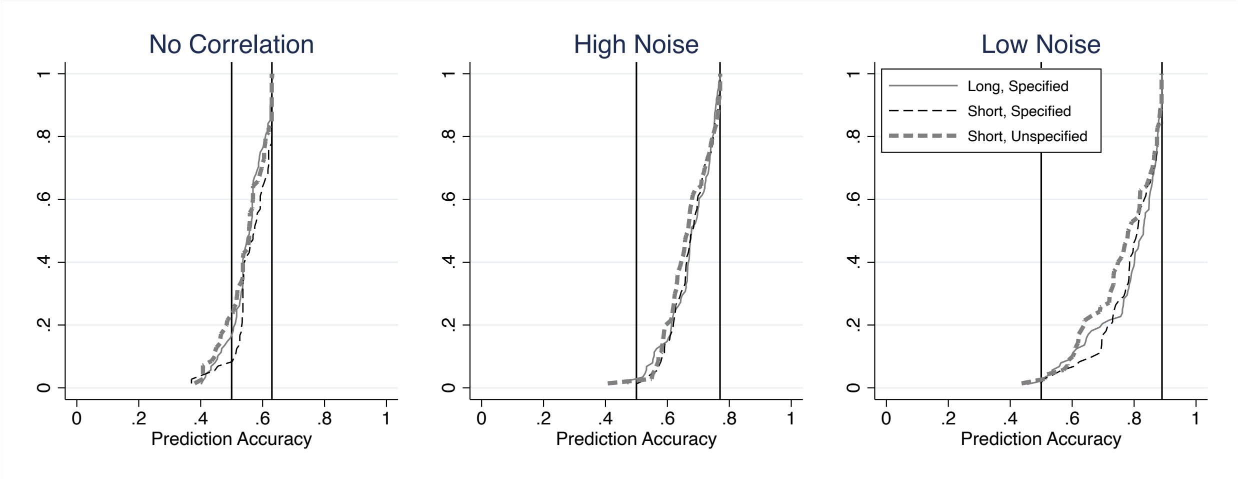


Figure 11: Distribution of Prediction Accuracy (CDF)

Notes: Long vs. Short refers to the length of the notes subjects were allowed to take for each data set. Explicit vs. Unspecified refers to how much information was provided to the subjects about the prediction task. See Section 3.2 for details. The two vertical lines denote the prediction probability for an agent who guesses randomly (on the left) vs. one who guesses optimally (on the right). Each observation takes the average prediction accuracy by subject in the data sets corresponding to the specific category. Equality of the distributions cannot be rejected by a Kolmogorov-Smirnov test (p value > 0.10 in all pairwise comparisons).

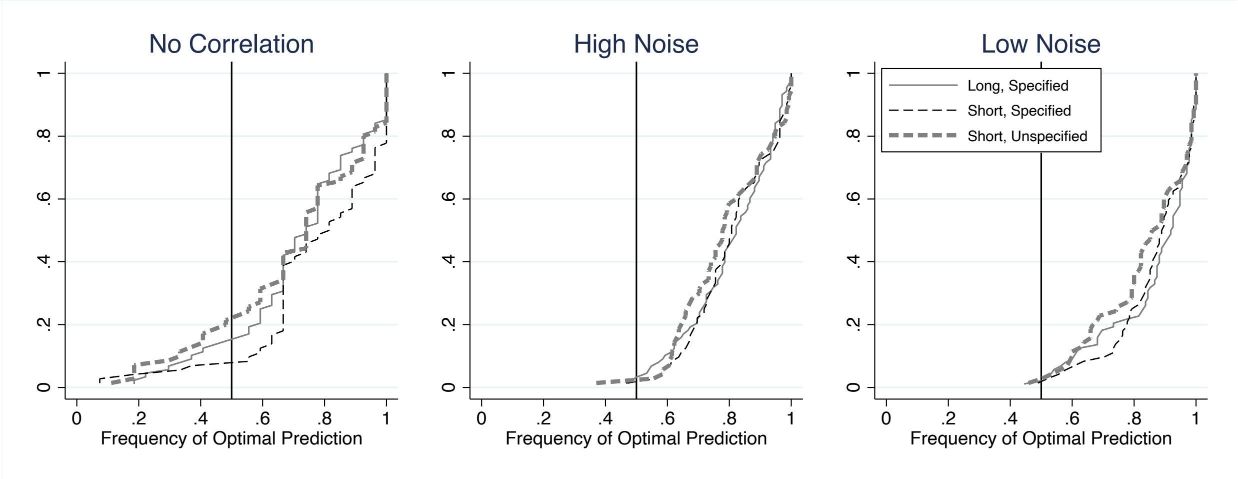


Figure 12: Distribution of Optimality of Guesses (CDF)

Notes: Long vs. Short refers to the length of the notes subjects were allowed to take on each data set. Explicit vs. Unspecified refers to how much information was provided to the subjects on the prediction task. See Section 3.2 for details. The vertical line denotes the frequency of making the optimal guess for an agent who guesses randomly. Equality of the distributions cannot be rejected by a Kolmogorov-Smirnov test (p value > 0.10 in all pairwise comparisons).

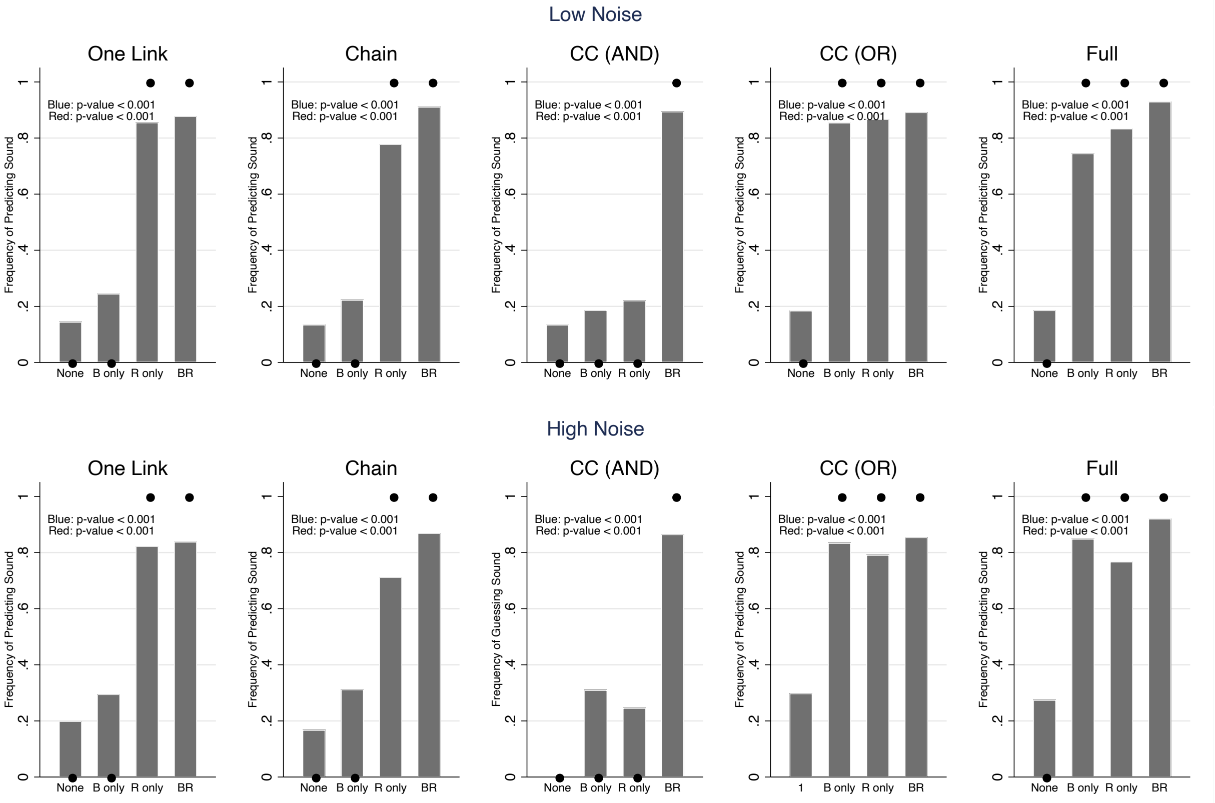


Figure 13: Guesses by Light Configuration

Notes: *None* refers to trials where both red and blue lights are off; *B only* refers to trials where only blue light is on; *R only* refers to trials where only red light is on; *BR* refers to trials where both lights are on. The first category (None) is missing for Common Consequence (AND) with high noise (see Online Appendix A for details on parametrization). Black dots denote optimal behavior.

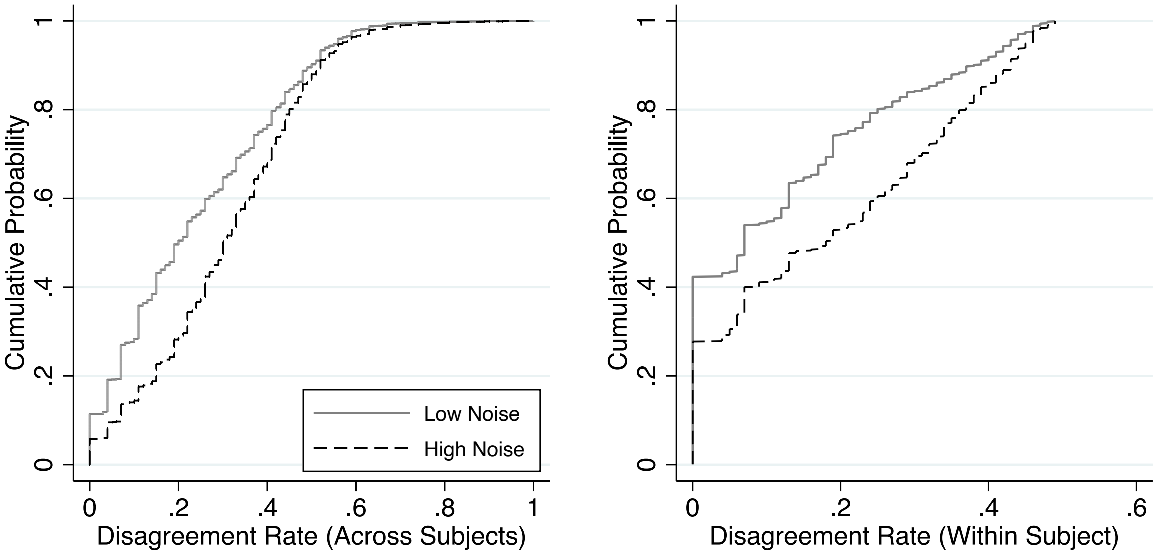


Figure 14: Disagreement Rate by Noise Level

Notes: Disagreement rate across subjects corresponds to the likelihood that any two participants make opposing predictions on the sound after observing the same information on the status of the other variables. This measure is computed at the dataset level for any pair of participants. Disagreement rate within subject corresponds to the likelihood that a participant makes opposing predictions on the sound after observing the same information on the status of the other variables. This measure is computed at the dataset level for each participant.

D Details and Robustness of Typing Procedure

D.1 Restricting Focus to Stationary Prediction Rules

In our analysis, we focus on a subset of prediction rules that only condition on the light configuration. Namely, we do not allow for prediction rules that condition on the order or sequencing of the dataset or the prediction task. Several measures were taken in the design of the computer interface to minimize the salience and attractiveness of such rules.

- Rows of the data set was deliberately *not* labeled from 1 to 27.
- All rows of the data set were presented simultaneously.
- Order of the rows were randomized on the individual level.
- Prediction tasks were presented one at a time, deliberately *not* labeled from 1 to 27.
- Order of the prediction tasks were also randomized on the individual level and differed from the order associated with the rows observed in the data set.

Despite these measures, in this section study whether there is any evidence pointing towards the use of such rules. First, the *notes* participants write about each data set are useful as they reveal what aspects of the data set they are attentive to. We do not find any evidence of direct reference to the order or sequencing of the data set (or some other form of serial correlation). But, as reported in more detail in Section **??**, a non-negligible share of participants code the data set in its entirety in the *notes*. We consider two ways in which participants might condition on the order of events in the data set: (1) Their predictions do not condition on the lights, but match the order with which the sound was observed to be on or off in the data set; (2) Their predictions condition on the lights, and for each light configuration, predictions match order with which the sound was observed to be on or off in the data set. We compute the share of participants whose *notes* code the order of events, and their predictions match behavior as described in (1) or (2) better than *any* of the 16 deterministic prediction rules considered in our analysis. This share is 3.7 percent. 88 percent of such observations are classified as not corresponding to a deterministic rule in our analysis. Among these observations, in only 5.8 percent of cases rules (1) or (2) achieve a fit (match predictions) at a rate higher than 90 percent. Overall, these results suggest that if some participants are drawn to using more complex prediction rules that take into account the order of events in the data set, there is limited scope for such behavior.

D.2 Estimation on the Population Level

We assume that there are 17 possible types. The first 16 are associated with the deterministic prediction rules described in Table 5.[57] Subjects following a deterministic rule are assumed to implement their rule with some noise where the likelihood of making a prediction consistent (inconsistent) with the rule is β $(1-\beta)$ where $\beta \in (0.5, 1]$. In addition, we allow for a stochastic type that predicts the sound to be on with probability $\delta \in [0, 1]$ independent of the light combination.

Let $p \in \Delta^{17}$ denote the share of each type in the population. Let $l \in \{nn, rn, nb, rb\}$ denote all possible light configurations in the first 27 predictions rounds of part 2 of the experiment where the status of both the red and blue lights are revealed to subjects (n denotes either the blue or red light to be off).

Now we describe the behavior of the subjects. Fix the data set. Let $g_{it} \in \{0, 1\}$ denote whether subject i guesses the sound to be on or off in round t of the prediction task. Let $y_{itr} \in \{0, 1\}$ denote whether this guess was consistent with deterministic prediction rule $r \in \{1, .., 16\}$ given the light configuration l observed in that round for the subject. Given p, β and δ, the log likelihood of observing subjects' predictions according to the model can be written as follows:

$$\sum_i \log \left(\sum_{r=1}^{16} p_r \prod_{t=1}^{27} \beta^{y_{itr}} (1-\beta)^{1-y_{itr}} + p_{17} \prod_{t=1}^{27} \delta^{g_{it}} (1-\delta)^{1-g_{it}} \right) \tag{1}$$

Parameters p, β and δ are estimated to maximize the log likelihood function stated in equation 1.

[57] We make one exceptions to this in the Common Consequence (AND) high noise data set, where not all light configurations are observed. In this case, we treat the few deterministic rules that are observationally equivalent as the same. If one of these happens to be one of the top five as listed in Table 5, we present it as this rule in reporting the results.

D.3 Results on the Aggregate Level

Table 7: Population Level Estimates for Prevalence of Different Prediction Rules by Data Set

	Deterministic Rules							Non Deterministic	
	G All	*G w/ R*	*G w/ B*	*G w/ R & B*	*G w/ R or B*	Other	β	Share	δ
One Link, Low Noise	0.02	**0.64**	0.01	0.01	0.06	0.03	0.95	0.23	0.57
One Link, High Noise	0.03	**0.56**	0.01	0.02	0.07	0.05	0.93	0.26	0.58
Chain, Low Noise	0.02	**0.61**	0.01	0.13	0.08	0.01	0.96	0.14	0.59
Chain, High Noise	0.02	**0.48**	0.03	0.14	0.12	0.02	0.93	0.20	0.60
CC (AND), Low Noise	0.03	0.04	0.01	**0.70**	0.02	0.02	0.96	0.17	0.60
CC (AND), High Noise	0.00	0.02	0.04	**0.62**	0.09	0.00	0.93	0.22	0.59
CC (OR), Low Noise	0.05	0.01	0.02	0.00	**0.69**	0.04	0.95	0.18	0.57
CC (OR), High Noise	0.04	0.00	0.02	0.01	**0.52**	0.02	0.94	0.39	0.62
Full, Low Noise	0.03	0.08	0.00	0.06	**0.64**	0.01	0.95	0.18	0.62
Full, High Noise	0.04	0.03	0.03	0.02	**0.54**	0.04	0.92	0.29	0.62

Notes: Table reports estimates for the mixture model. For example, with the One Link, Low Noise data set 64 percent of subjects are classified as predicting the sound to be on only when the red light is on. The optimal rule for each data set is highlighted in bold. See Table 5 in Online Appendix B for descriptions of each prediction rule.

D.4 Typing Procedure

We use mixture model estimates to type subjects on the individual level. Namely, using mixture model estimates as a prior, given the guessing patterns of each subject for each data set, we estimate the posterior probability they are following each of the 17 rules described above. We then classify each subject as following the rule with highest posterior probability.

To demonstrate this, consider a simpler setup where the set of candidate rules consist of only two: *G w/ R* and *G w/ B*, where the former (latter) rule predicts $S = 1$ only when $R = 1$ ($B = 1$). Let c_r^i denote the number of rounds (out of 27) in which participant i's guesses are consistent with rule r and p_r denote the population level estimate for the prevalence of this rule. If implementation error rate is estimated to be ϵ, the posterior probability participant i is using rule r would be computed as follows: $\frac{p_r(1-\epsilon)^{c_r^i}\epsilon^{27-c_r^i}}{\sum\limits_{k\in\{1,2\}} p_{r_k}(1-\epsilon)^{c_{r_k}^i}\epsilon^{27-c_{r_k}^i}}$.

D.5 Results on the Individual Level

Table 8: Type Shares for Different Prediction Rules by Data Set

	Deterministic Rules					Non Deterministic	
	G All	*G w/ R*	*G w/ B*	*G w/ R & B*	*G w/ R or B*	β	Share
One Link, Low Noise	0.02	**0.64**	0.01	0.01	0.07	0.03	0.23
One Link, High Noise	0.03	**0.57**	0.00	0.02	0.06	0.05	0.27
Chain, Low Noise	0.02	**0.59**	0.01	0.15	0.07	0.01	0.15
Chain, High Noise	0.02	**0.47**	0.03	0.17	0.10	0.02	0.19
CC (AND), Low Noise	0.03	0.04	0.00	**0.71**	0.02	0.02	0.17
CC (AND), High Noise	0.00	0.02	0.03	**0.65**	0.09	0.00	0.21
CC (OR), Low Noise	0.05	0.01	0.02	0.00	**0.70**	0.04	0.18
CC (OR), High Noise	0.03	0.00	0.02	0.01	**0.51**	0.03	0.40
Full, Low Noise	0.03	0.10	0.00	0.06	**0.62**	0.01	0.18
Full, High Noise	0.03	0.02	0.02	0.01	**0.62**	0.03	0.27

Notes: Table reports share of participants classified as using each prediction rule for each data set. For example, with the One Link, Low Noise data set 64 percent of subjects are classified as predicting the sound to be on only when the red light is on. The optimal rule for each data set is highlighted in bold. See Table 5 in Online Appendix B for descriptions of each prediction rule.

D.6 Simulation Results

The goal of this section is to demonstrate that type shares in our experiment can reliably be recovered for each data set following the typing procedure described above. To do this, for each data set we take the estimated type shares (as well as implementation error and mixing probability for the stochastic type) estimated in the paper as in input and we simulate behavior (predictions about the sound given different light configurations) in 1000 times. We then use our typing procedure exactly as described above, which involves first estimating a mixture model and then using these as a prior to type participants, to estimate type shares.

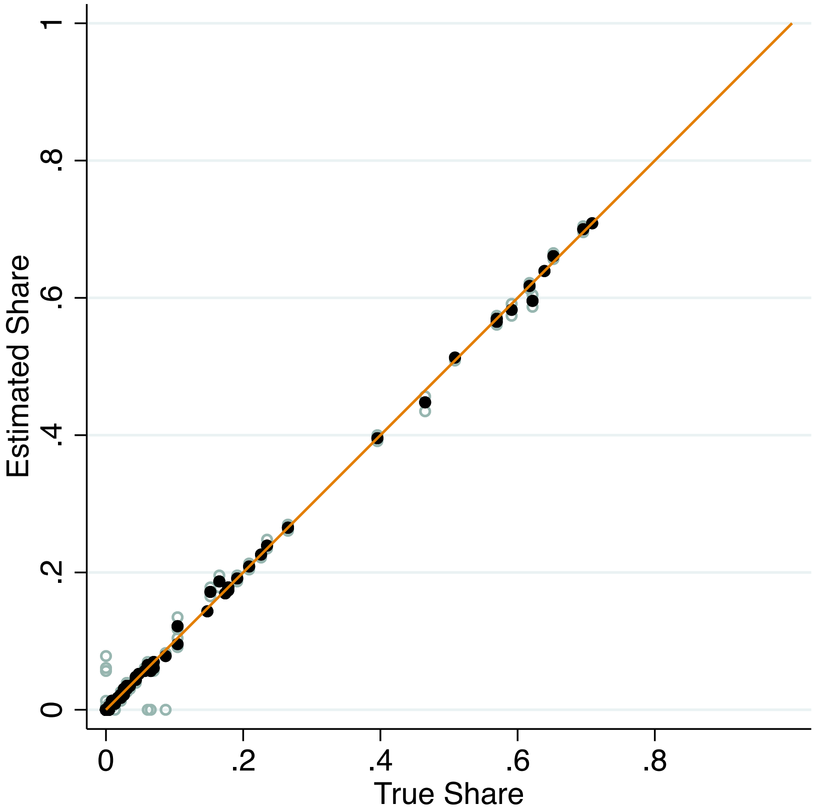


Figure 15: Estimation results using simulations

Notes: Estimation results from 1000 simulated experiments with 230 subjects for 10 different data sets. *True Share* refers to input values to the simulations which correspond to estimates reported in the paper. Solid dots represent median estimate, hollow bubbles represent 25th and 75th percentile estimates.

E Further Analysis on the Individual Level

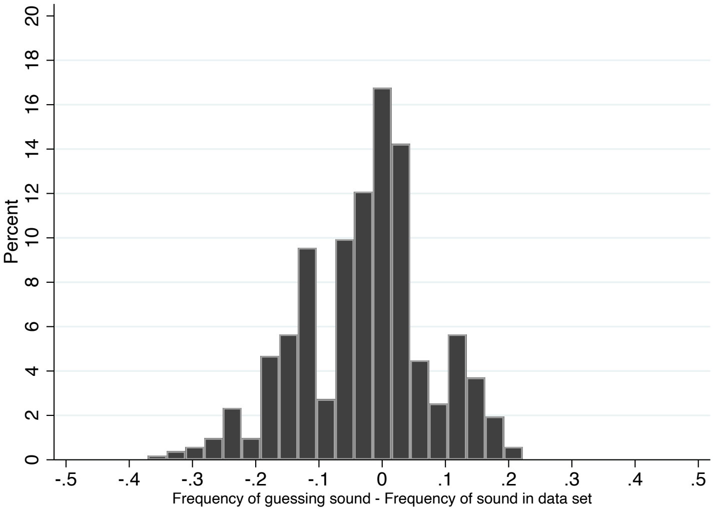


Figure 16: Distribution of Distance between Frequency of Predicting the Sound and the Frequency of the Sound in the Data Set

Notes: Includes only participant-data set observations which are typed to correspond the stochastic rule.

Table 9: Marginal Effects (Probit) of Between Subject Treatment Variation on Behavior in One Link and Chain

	Optinal	Misalignment error	Non-conditioning error
Unknown Prediction	-0.269*	0.209	0.163
	(0.152)	(0.154)	(0.173)
Short Notes	-0.0831	0.281*	-0.149
	(0.151)	(0.145)	(0.177)
Observations	920	920	920

Controls for each data set are not reported.

Standard errors (clustered at the subject level) in parentheses.

***1%, **5%, *10% significance.

Table 10: Marginal Effects (Probit) of Between Subject Treatment Variation on Behavior in Common Consequence and Full

	Optinal	Misalignment error	Non-conditioning error
Unknown Prediction	-0.102	0.0809	0.0796
	(0.143)	(0.137)	(0.157)
Short Notes	0.0479	0.118	-0.109
	(0.136)	(0.123)	(0.152)
Observations	1380	1380	1380

Controls for each data set are not reported.

Standard errors (clustered at the subject level) in parentheses.

***1%, **5%, *10% significance.

Computing Distance from Bayesian Behavior

Consider an agent who is uncertain about the likelihood of an event p. In our experiment, we p can denote the likelihood with which $S = 1$ conditional on (R, B). The agent's prior is given by the Beta distribution and is characterized by two parameters p_0 and η, such that:

$$\mathbb{E}(p \,|\, p_0, \eta) = p_0 \quad \text{and} \quad \mathbb{V}(p \,|\, p_0, \eta) = \frac{p_0(1 - p_0)}{\eta + 1}.$$

While p_0 denotes the expected value of p, η captures the strength of the prior and, hence, can be interpreted as a measure of the agent's confidence.[58]

The agent updates beliefs on p using outcomes from a Bernoulli process where the probability of the event happening is the true p. The data observed by the agent can be characterized by two parameters: the number of observations n, and the observed frequency of the event among these observations f. The agent's updated posterior is still characterized by a Beta distribution with adjusted parameters $\tilde{p}$ and $\tilde{\eta}$:

$$\tilde{p} = \left(\frac{\eta}{\eta + n}\right) p_0 + \left(1 - \frac{\eta}{\eta + n}\right) f \tag{2}$$

In summary, the model describes how beliefs evolve with feedback as a function of two parameters: p_0, prior expected value on p; and η, a measure of initial confidence.

Note that it is optimal to predict $S = 1$ when $\tilde{p} \geq 0.5$. This happens when

$$n(f - 0.5) > \eta(0.5 - p_0) \tag{3}$$

While we do not observe the agent's prior p_0 and η in the experiment, we observe $n(f - 0.5)$. Let's refer to $n(f - 0.5)$ as the *strength of evidence*. One important implication of this simple model is that the agent's prediction must be monotonic in $n(f - 0.5)$, the strength of evidence. That is, if the agent faces distinct data sets with different features, the optimal prediction for the agent (conditional on a light configuration) will be $S = 0$ for values of $n(f - 0.5) < \bar{e}$ for some $\bar{e} := \eta(0.5 - p_0)$ and $S = 1$ for values of $n(f - 0.5) > \bar{e}$. While we do not know $\bar{e}$, we can search for whether there exists a value of $\bar{e}$ that rationalizes the agent's decisions.

Figure 17 depicts the behavior of four different participants to display variation in behavior. Participant 209 is perfectly optimal, consistently predicting $S = 1$ only when the evidence in the

[58] In the standard formulation, the Beta distribution is characterized by two parameters: α, β such that $\mathbb{E}(p \,|\, \alpha, \beta) = \frac{\alpha}{\alpha+\beta}$ and $\mathbb{V}(p \,|\, \alpha, \beta) = \frac{\alpha\beta}{(\alpha+\beta)^2(1+\alpha+\beta)}$. The mapping to p_0 and η are such that $p_0 = \frac{\alpha}{\alpha+\beta}$ and $\eta = \alpha + \beta$.

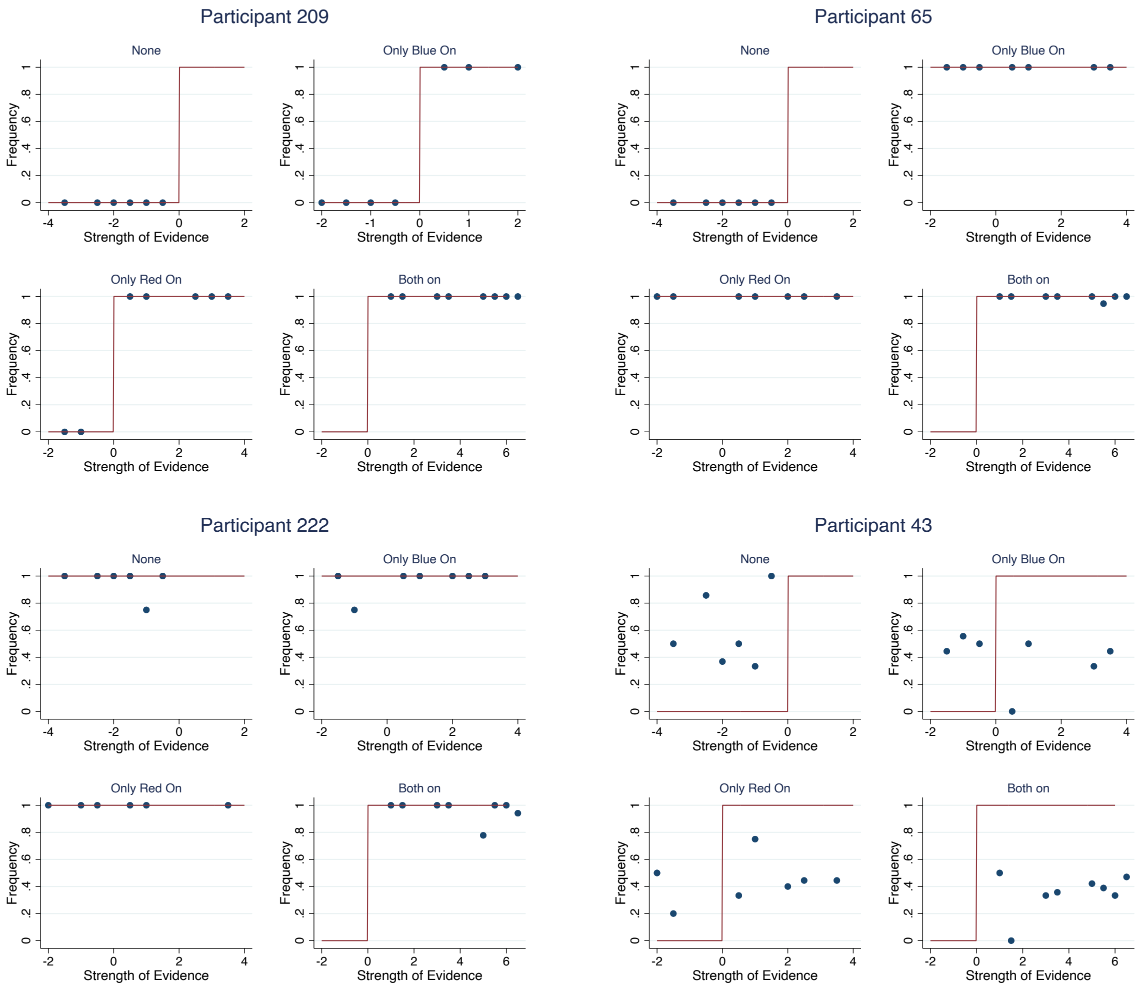


Figure 17: Examples of Participant Behavior

Notes: Examples on how participants predictions (frequency of guessing $S = 1$ changes with the strength of evidence. Red lines for participants 209, 65 and 222 depict how behavior can be rationalized as Bayesian with respect to some p_0 and η.

data points in that direction. Participant 65's predictions perfectly follow *G w/ R or B* in all data sets independent of the strength of evidence. But this can be rationalized with a strong prior as depicted with the red lines. This participant would be classified as conditioning on an irrelevant variable (displaying a misalignment error) in some of the data sets. Participant 222's predictions very closely follow *G All* in all data sets independent of the strength of evidence. But this can be rationalized with a strong prior as depicted with the red lines. This participant would be classified as ignoring a relevant variable (displaying non-conditioning error) in some of the data sets. Participant 43, in contrast to the other three examples, displays highly stochastic behavior. It is easy to see why such behavior cannot be rationalized as a Bayesian response to some prior.

The examples provided in Figure 17 depict the spectrum of behavior observed in our experiment. While behavior of participants 209 and 65 are different, they can each be perfectly rationalized as Bayesian behavior with respect to some prior. For participant 222, we see some small deviations from Bayesian behavior. One way to see this is that, only a few predictions of this participant would need to be modified (flipped) for them to be perfectly consistent with Bayesian behavior. Participant 43's predictions are far from Bayesian; namely, many predictions of this participant would need to be changed to reconcile with Bayesian behavior.

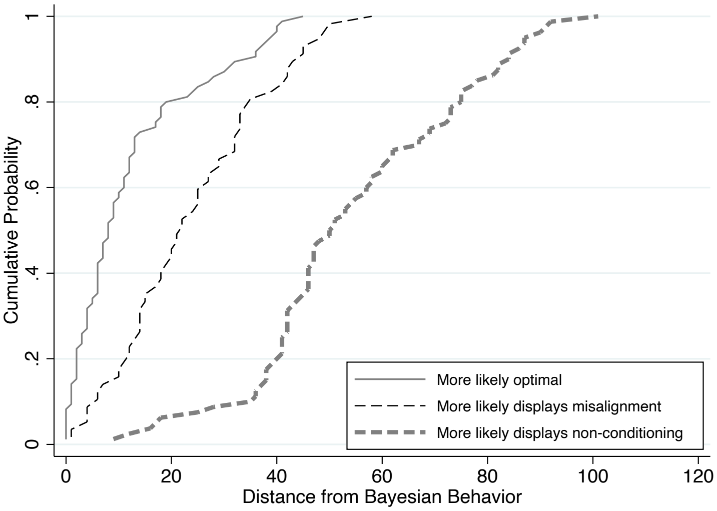


Figure 18: Distribution of Distance to Bayesian Behavior

Notes: Focusing on One Link and Chain data sets we classify participants into three categories: Most likely to be classified as using the optimal prediction rule, most likely to be classified as ignoring a relevant variable, and most likely to condition on an irrelevant one. Distance to the Bayesian Behavior measure is computed over all ten data sets.

Building on this observation, we define Δ_i, *distance to Bayesian behavior* for agent i, as minimum number of predictions (among 27×10 observed for agent i) that would need to be flipped such that the behavior of the agent can be rationalized with some value of $\bar{e}$ as described before. Figure 18 plots the distribution of this measure for three group of participants. We observe that very few subjects behave in a way that is perfectly aligned with Bayesian behavior. Nonetheless, the third group of participants, those who are more likely to display non-conditioning errors (ignore all relevant variables) appear to be very different from the other two. Namely, it is much more difficult to reconcile their behavior as a Bayesian response to some prior. This is consistent with the example of Participant 43 as depicted in Figure 17. Overall, we find that ignoring all relevant variables is often associated with stochastic behavior in our data.

Behavior With Partial Information on Lights

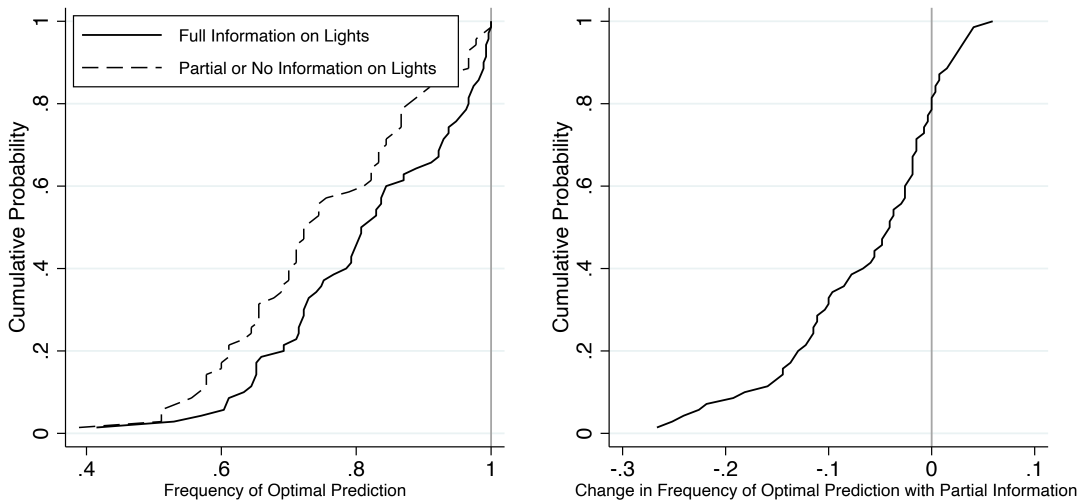


Figure 19: Behavior with Partial Information on Lights

Notes: Left panel plots distribution of frequency of optimal prediction contrasting rounds 1-27 (where there is always full information on the status of the lights) with rounds 28-36 (where there is partial or no information on the status of the lights). Right panel plots the distribution of decline in frequency of optimal predictions when comparing the first 27 rounds to the last nine rounds. In both graphs a unit of observation is a participant's behavior in a single data set.

Examples on Classification of Notes

Codes All Data

srb nrb srb n srb*4 n srb nrb nrnr srb*3 nr n nrb srb n srb sb nrb srb*2 sr
$|BD|N|A|D|N|RD|BD|RD|N|D|N|BD|BD|N|N|RD|BD|BD|N|D|A|N|D|R|BD|D|B|$

Summarizes Frequency

b+s=1 b only=3 r+s=7 r only=2 r+b+s=8 r+b only=2 none+s=1 none=3

12/27 Both - 12D 13/27 none - 3D 1/27 R only - 1D 1/27 B only - 1D

Identifies correlations

both on sound, B no sound, R no sound

ding when both are on

Does not identify correlations

blue

Many dings, many all three

Notes and Mistakes

Table 11: Notes and Mistakes in Common Consequence AND/OR and Full

	Share	Deterministic	Optimal (no errors)	Optimal (w/ errors)	Ignores (only one)	Ignores (both)
Codes All Data	.21	.69	.19	.41	.06	.34
Summarizes Frequency	.25	.92	.50	.37	.02	.10
Identifies Correlations	.28	.82	.31	.33	.15	.21
Other	.27	.63	.15	.28	.10	.46

Notes: See notes for Table 2.

Table 12: Behavior in Common Consequence and Full

		High Noise			
		Optimal (no errors)	Optimal (w/ errors)	Non-conditioning error (Ignore relevant)	Misalignment error (Cond. suboptimally)
Low Noise	Optimal (no errors)	**.52**	.35	.08	.04
	Optimal (w/ errors)	.12	**.46**	.36	.06
	Non-Conditioning error	.04	.18	**.70**	.07
	Misalignment error	.16	.37	.39	**.12**

Notes: The table reports likelihood of different categories of behavior in the high noise data set of each DAG as a function of category of behavior in the low noise data set of the same DAG. Example: 70 percent of subjects who ignored both relevant variables in CC (AND) L, CC (OR) L or Full L also ignore both relevant variablse in CC (AND) H, CC (OR) H or Full H.

Response Time

Table 13: Determinants of Prediction Accuracy and Optimality (OLS)

	Prediction Accuracy	Prediction Optimality
Time spent taking notes	-0.000706	-0.000961
	(0.00109)	(0.00171)
Time spent making predictions	0.00938***	0.0141***
	(0.00242)	(0.00383)
Observations	2300	2300

Controls for each data set are not reported.

Standard errors (clustered at the subject level) in parentheses.

***1%, **5%, *10% significance.

Table 14: Average Response Time

	All Data		Treatments with Short Notes	
	Time taking notes	Time making predictions	Time taking notes	Time making predictions
Codes All Data	3.21	2.81	3.55	2.70
Summarizes Frequency	3.49	2.25	3.32	1.75
Identifies Correlations	2.37	1.84	2.11	1.54
Other	3.08	2.10	3.55	1.77

Notes: Reported values are in minutes. Includes all data sets where optimal behavior involves conditioning on the lights.

Learning

Table 15: Determinants of Prediction Accuracy and Optimality (OLS)

	Prediction Accuracy	Prediction Optimality
Case	-0.00105*	-0.00189*
	(0.000574)	(0.000980)
Observations	2530	2530

Case (from 1 to 11) reflects the order in which data set was observed.

Controls for each data set are not reported.

Standard errors (clustered at the subject level) in parentheses.

***1%, **5%, *10% significance.

Table 16: Marginal Effects (Probit) on Determinants of Conditioning on a Light

Optimal to Cond. on Light	1.038***
	(0.0753)
Optimal to Cond. on Opposite Light	-0.506***
	(0.0574)
Optimal to Cond. on Same Light Previous Case	0.0750**
	(0.0366)
Optimal to Cond. on Opposite Light Previous Case	-0.0374
	(0.0335)
Case	-0.0173**
	(0.00722)
Observations	4600

Case (from 1 to 11) reflects the order in which data set was observed.

Controls for each data set are not reported.

Standard errors (clustered at the subject level) in parentheses.

***1%, **5%, *10% significance.

Disagreement Rates

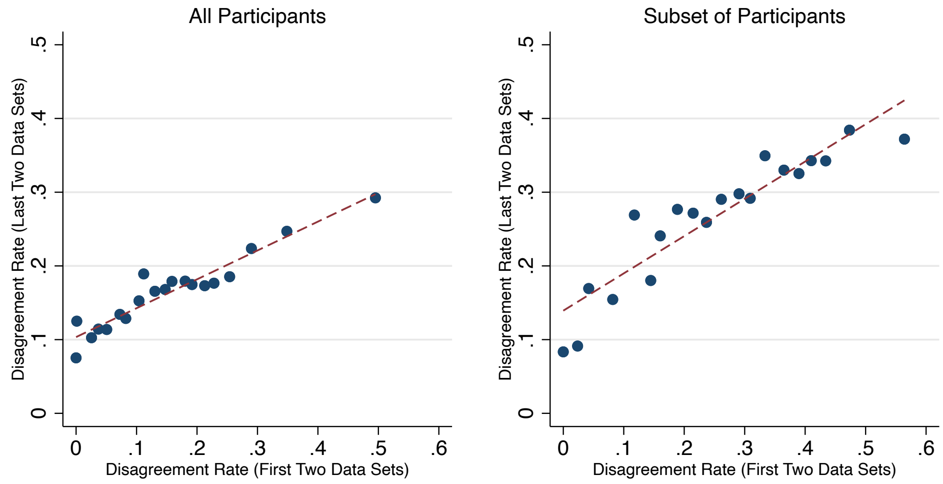


Figure 20: Disagreement Rates in Early and Late Data Sets Among Participants Classified as Using a Deterministic Prediction Rule

Notes: See Figure 14.

F Protocol for Coding Participants' Notes

Two undergraduate research assistants were provided with a spreadsheet that contained the notes written by participants all treatments. Research assistants were not informed of any aspect of the research beyond the information in the spreadsheet, the instructions to the experiment, and the coding protocol that we attach in this appendix. The coding protocol describes the spreadsheet that coders were provided with as well as the questions that they were asked to answer. Both assistants were asked to code all notes independently. For any differences in coding, they were asked to have a meeting, discuss each difference and reconcile them. We reproduce the protocol next. The screenshot referred to in the protocol corresponds to Figure 1.

Protocol for coding notes

You will code notes from an experiment. The purpose of this protocol is to explain you how to do the coding. Before describing your task, we will explain how the notes that you will code were generated. A participant in the experiment that generated the notes was provided with data from 11 machines. Each machine produced 27 trials. The screenshot on this page provides an example for machine 1. In every trial the machine makes a sound or does not make a sound (last column). A machine has lights of multiple colors, including always one red and one blue light. In every trial, the participant can see whether the red light is on or off (first column), whether the blue light is on or off (second column), but cannot see other lights, which can be on or off (third column). Their task in a later part of the experiment is to make predictions. They will see the status of the red and blue light and their task is to predict whether the machine will make a sound or not. When they face the prediction task, the data that you see in the screenshot below is NOT available to them. But in this first part of the experiment, when they do see the data, they can write notes on the left side of their screens. These notes will be available when they face the prediction task. Each participant wrote notes for themselves on each of the 11 machines that they faced. Your task is to read each note and determine whether the note has certain properties that we will describe later.

Make yourself familiar with the spreadsheet (textdata_tocode.xls). The first column (column A) is the treatment. In Treatments 1 and 2, the task the prediction task consists of we providing them with the status of the red and blue lights and they having to predict whether the machine makes a sound or not. The difference is that in Treatment 1 notes can be up to 280 characters long. In Treatment 2 notes can be up to 75 characters long. In Treatment 3, the prediction task is broader: they are given partial information about a trial (e.g., one light) and are asked to predict information that they are not provided (e.g., the other light or the sound). In Treatment 3, notes are 75 characters long. What treatment the participant is, does not change the questions you have to have to answer to code each note. The second column, Column B, is the number id that a specific participant was given. The third column, Column C, captures the specific machine the note corresponds to. There are 11 machines or cases, so this variable goes from 1 to 11. Column D (dagnotes) contains the written note that you have to code. For each row, you will fill in several columns, starting from the fifth column onward. If a note is blank (the participant did not write down anything), move on to the next note. For non-blank notes, you will fill in Columns E to H:

E. FullData: Did the participant code the full 27 rows of data in their own words (for example, using their own coding)? [Note: it does not matter if you can or cannot make sense of the way in which they coded the data. All you have to assess is if the participant seemed to have used an approach such that when they read their own notes, they could use their notes to reconstruct the 27 data points.]

- Enter 1: if the participant's notes would let you reconstruct the order in which the data points were presented and the participant does not try to use any code. Example, the participant would use something like 'Red On, Blue On, Sound On', for a trial in which that is what happens. Or a small modification like 'R On, B On, S On.' And the message would then have 27 instances in which the full event is written. If the notes attempt to do this but there is a mistake (for example because they record 26 instead of the 27 trials), enter 1.5 instead of 1.

- Enter 2: if the participants notes would let you reconstruct the order in which the data points were presented, and the participant uses a code to describe the 8 possible cases in the trials. Example, someone uses letter 'A' for the case Red On, Blue On, Sound On; letter 'a' for the case Red On, Blue On, Sound Off; uses letter 'B' for the case Red On, Blue Off, Sound On; letter 'b' for the case Red On, Blue Off, Sound Off; uses letter 'C' for the case Red Off, Blue On, Sound On; letter 'c' for the case Red Off, Blue On, Sound Off; uses letter 'D' for the case Red Off, Blue Off, Sound On; letter 'd' for the case Red Off, Blue Off, Sound Off. Then, the message would involve 27 characters: d c a A D B b a C d c a A D D d a C d c a a c A b a C. This is an example in which the way that the participant coded the message fully lets you reproduce the data set, including the order in which the trials appeared, and the participant uses code for each possible case. If the notes attempt to do this but there is a mistake (for example because they record 26 instead of the 27 trials), enter 2.5 instead of 2.

- Enter 3: if the participants message uses a code but would not let you reproduce the order in which the trials appeared. Following the previous example, it could be something like: A 4/27; a 3/27, and so on. If the notes attempt to do this but there is a mistake (for example because they divided by 26 instead of 27), enter 3.5 instead of 3.

- Enter 4 if the participant uses numerical information that would not let you fully reproduce the data set. For instance, the participant just says that the sound takes place 14/27 times. If the notes attempt to do this but there is a mistake (for example because they divided by 26 instead of 27), enter 4.5 instead of 4.
- Enter 99: if you are unsure.
- If the answer is no, leave blank.

F. Generic: Do the notes use words to describe the data?

- Enter 1: if the note mentions that the sound is random or something similar. [For example, 'the sound is sometimes on, sometimes off, for certain light combinations.']
- Enter 2: if the note mentions that the sound is 'mostly' or 'overall' on without specifying the frequency or connecting it to a combination of lights. [For example, 'in most of the trials the sound is on.']
- Enter 3: if the note mentions that the sound is always on or off for certain light combinations. [For example, the sound is always on when the blue light and the red light are on.]
- Enter 4: if the note uses mostly words but cannot be classified as any of the previous three. Here is an example of a note that would fall in this category. 'Mach 8: Intervals of 5x ding followed by 1-2 silence. As more trials done, more ding. Seemingly no relation between r/b light and ding. Most r&b=n = no ding. Most r&b=y = ding. There are exceptions.' It does not use the word 'random' directly, so it is not captured by '1.' It does mention a word like 'most' but it also connects the sound to the lights, so it does not fall into '2.' It does not convey that the light is on or off always for some combination of lights, so it does not fall into '3.' But while there are some numbers in this note, it involves mostly words. Since it does not fall into 1, 2 or 3 and it does use mostly words, it qualifies as 4.
- If none of the above aspects is mentioned, leave blank.

G. Correlation: Does the participant uses the word correlation, association, or a related word to describe the data?

- Enter 1: if the answer is yes.
- Enter 99: if you are unsure.
- If the answer is no, leave blank and jump to column F.

H. Causality: Does the participant use the word ‘causality’ or the idea that one item causes another item? [Example: ‘When the blue light is on, the machine makes a sound’ or ‘When the machine makes a sound, then the red and the blue lights are on’] Does the participant use the word ‘lack of causality’ or the idea that one item does not cause another item? [Example: ‘When the blue light is on, the machine makes a sound’ or ‘When the machine makes a sound, then the red and the blue lights are on’]

- Enter 1: if the answer to either question above is yes.
- Enter 99: if you are unsure.
- If the answer is no, leave blank and jump to column G.

I. Model [Answer this question only if you did not answer E as 1, 2, 3 or 4.] Assume you have to make predictions, as participants in this experiment did. Looking at the notes they wrote, evaluate if any of the following is a good match to what the notes suggest as best strategy for predictions for when the sound is on:

- Enter 0: if the notes suggest predicting that the machine makes the sound all the time.
- Enter 1: if the notes suggest predicting that the machine makes the sound only when the blue light is on.
- Enter 2: if the notes suggest predicting that the machine makes the sound only when the red light is on.
- Enter 3: if the notes suggest predicting that the machine makes a sound only when either the red light or the blue light is on.
- Enter 4: if the notes suggest predicting that the machine makes a sound only when both the red light and the blue lights are on.
- Enter 5: if the notes suggest a clear prediction strategy that is not one of the above (example: notes suggest predicting that the machine makes the sound only when the blue light is off).

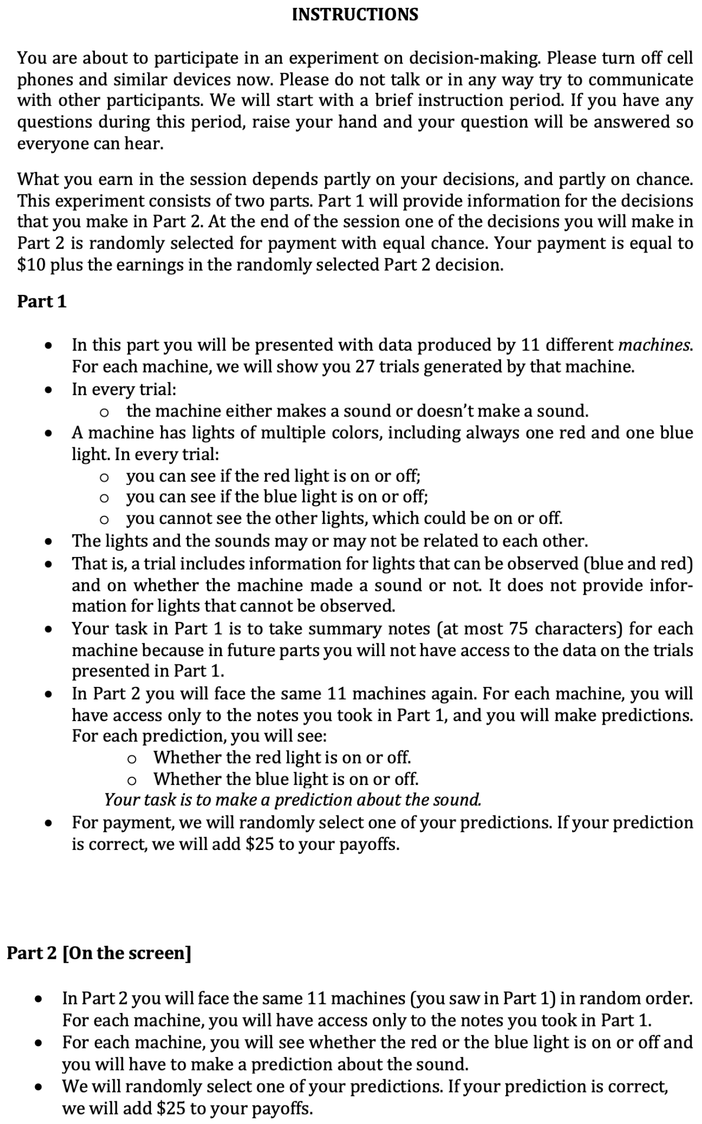
- Enter 6: the notes clearly do not suggest a prediction strategy.
- Enter 99: if you are unsure. If you dont know whether it is possible to classify as any of the first 5, enter 99. If you think it may be one of the five but are unsure, do the following. If you think it may be classified as 2, then enter 99.2. If you think it may be classified as 3, enter 99.3. Example: If the note says, Sound is on when red light is on but not otherwise, then you would clearly code that as 2. But if the note says: Red on, sound on 20/27; when Red off, sound not there very often, it may be unclear because the note does not specify what to do if the red light is off. In this case, you may code it as 99.2.

J. Other aspects:

- Enter 1: if the note is very difficult to understand (incomprehensible).
- Enter 2: if the note compares this machine to previous machines.
- Enter 3: if the note conveys that there is a time dependency on trials (e.g., for a given machine, the participant thinks that after a trial in which the blue light was on and there was a sound, there will be another trial in which the red light will be on and there will not be a sound).

G Instructions

The followings are instructions from the Explicit Treatment.



INSTRUCTIONS

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

What you earn in the session depends partly on your decisions, and partly on chance. This experiment consists of two parts. Part 1 will provide information for the decisions that you make in Part 2. At the end of the session one of the decisions you will make in Part 2 is randomly selected for payment with equal chance. Your payment is equal to $10 plus the earnings in the randomly selected Part 2 decision.

Part 1

- In this part you will be presented with data produced by 11 different *machines*. For each machine, we will show you 27 trials generated by that machine.
- In every trial:
  - the machine either makes a sound or doesn't make a sound.
- A machine has lights of multiple colors, including always one red and one blue light. In every trial:
  - you can see if the red light is on or off;
  - you can see if the blue light is on or off;
  - you cannot see the other lights, which could be on or off.
- The lights and the sounds may or may not be related to each other.
- That is, a trial includes information for lights that can be observed (blue and red) and on whether the machine made a sound or not. It does not provide information for lights that cannot be observed.
- Your task in Part 1 is to take summary notes (at most 75 characters) for each machine because in future parts you will not have access to the data on the trials presented in Part 1.
- In Part 2 you will face the same 11 machines again. For each machine, you will have access only to the notes you took in Part 1, and you will make predictions. For each prediction, you will see:
  - Whether the red light is on or off.
  - Whether the blue light is on or off.

  Your task is to make a prediction about the sound.
- For payment, we will randomly select one of your predictions. If your prediction is correct, we will add $25 to your payoffs.

Part 2 [On the screen]

- In Part 2 you will face the same 11 machines (you saw in Part 1) in random order. For each machine, you will have access only to the notes you took in Part 1.
- For each machine, you will see whether the red or the blue light is on or off and you will have to make a prediction about the sound.
- We will randomly select one of your predictions. If your prediction is correct, we will add $25 to your payoffs.