A Practical Guide to Updating Beliefs from Contradictory Evidence*

Evan Sadler†

September 3, 2020

Abstract

We often make high stakes choices based on complex information that we have no way to verify. Careful Bayesian reasoning—assessing every reason why a claim could be false or misleading—is not feasible, so we necessarily act on faith: we trust certain sources and treat claims as if they were direct observations of payoff relevant events. This creates a challenge when trusted sources conflict: Practically speaking, is there a principled way to update beliefs in response to contradictory claims? I propose a model of belief formation along with several updating axioms. An impossibility theorem shows there is no obvious best answer, while a representation theorem delineates the boundary of what is possible.

1 Introduction

People routinely make high stakes decisions based on complex information they have no way to verify for themselves. Consider these examples:

1. We all make choices about how to care for our health, including choices about whether to get vaccinated, how to treat ailments, how much to exercise, and what kind of food to eat. These choices depend on our beliefs regarding many questions

*I am especially grateful to David Hirshleifer and Ben Golub, and to participants at the 2017 Retreat on Information, Networks, and Social Economics, for invaluable feedback on this project. I am also grateful to Roy Radner, whose guidance inspired me to think about these issues. Numerous conversations with friends and colleagues have challenged me, helping to refine my interpretation of this model and many subtle points in the paper. In particular, I would like to thank Daron Acemoglu, Sarah Auster, Jean-Pierre Benoit, Dhruva Bhaskar, Yeon-Koo Che, Mark Dean, Rohan Dutta, Itzhak Gilboa, Barton Lipman, Elliot Lipnowski, Doron Ravid, Debraj Ray, Ariel Rubinstein, and Yuval Salant as well as seminar participants at Columbia, NYU, and London Business School. Finally, I would like to thank Joel Sobel and four anonymous referees for their incredibly thoughtful comments and their patience as I revised this paper. A previous version of this paper circulated under the title "Falsehoods."

†Columbia University – es3668@columbia.edu

of fact. Do vaccines cause autism or other negative side effects? Which if any alternative medicines are effective? Are genetically modified foods safe? Do fad diets have desirable effects? Behind each question is a large body of research that few if any of us will fully evaluate. Instead, we rely on doctors, government agencies, or other trusted sources to help us understand the consequences of different actions.

2. People decide what politicians to support—not just through voting, but also through donations and volunteer work—at least partly based on factual beliefs. Did Iraq possess WMDs prior to the 2003 invasion? Did the ACA include a provision instituting "death panels?" Do humans contribute to climate change? Did a politician engage in illegal activity? Did a politician support a particular policy in the past? Though one can verify the answers in principle, the cost of effort is typically prohibitive. Instead, we rely on what we hear through news reports and social interactions.

3. Many regulatory and legal decisions depend on an evaluation of scientific evidence. Does a particular chemical cause cancer in humans, and how large is the risk? Is a drug test accurate? Does DNA obtained from a crime scene match a suspect? Was evidence tainted or falsified? Agency heads, prosecutors, and judges rarely understand all of the underlying science, and they certainly do not gather and analyze data themselves. Instead, they rely on scientists and technicians who conduct studies and write reports.

In each example, someone must make a choice without fully understanding the information generating process. Source reliability is ambiguous, feedback is limited, and contradictory claims are common. Due to cognitive or other costs, direct verification is infeasible, so individuals take certain claims at face value—we typically have faith that our friends, doctors, news anchors, and lab technicians are not trying to mislead us. This necessity makes us vulnerable to false beliefs. Indeed in each of these examples, even sophisticated individuals often act on false beliefs. For instance, concern about the safety of GMO foods is more prevalent among the highly educated (Funk and Kennedy, 2016), the U.S. Congress voted to authorize the 2003 invasion of Iraq based on false intelligence, studies based on fabricated data can influence a scientific field for years (Brainard and You, 2018), and falsified lab tests have led to hundreds if not thousands of criminal convictions (Mettler, 2017).

Sharing falsehoods has never been easier, and the frequency with which we get exposed to them makes it increasingly difficult to separate the true and useful from the false and harmful. In principle, one should form a prior over all possible reasons that every person and every source might make a particular claim, update to a posterior based on observed events, and choose an action to maximize expected utility. In practice, one cannot readily form beliefs about the provenance of every scientific paper, news report, or tweet. This paper addresses the following question: Practically speaking, is there a principled way to update beliefs in response to contradictory claims?

To address this question, I introduce a model of belief formation that both limits the complexity of an agent's reasoning and accommodates belief in falsehoods. There is a finite set of states of the world—this is the agent's model. A proposition is a subset of states, what we normally call an event, that need not contain the true state. An agent sequentially encounters propositions and forms beliefs, which are simply a set of propositions held in memory. I interpret belief in a proposition as the belief that it contains the true state. On encountering a new proposition, the agent applies an update rule to arrive at new beliefs. Crucially, the update rule must allow the agent to process *any* proposition, whether or not it is consistent with her current beliefs. To rephrase my question: Are there "good" update rules in this setting?

I propose two families of axioms. "Non-manipulability" axioms capture the idea that a good update rule yields the same beliefs given the same information, no matter how that information is presented. In contrast, "willingness-to-learn" axioms capture the idea that a good update rule allows an agent to learn about the world. An impossibility theorem shows that no rule can simultaneously satisfy key axioms from both families: the ability to learn necessarily opens the agent to manipulation. I subsequently provide an axiomatic characterization of update rules, some of which display common psychological biases such as confirmation bias and motivated reasoning.[1] Each rule encompasses a particular constellation of non-manipulability and willingness-to-learn axioms. In light of the impossibility result, this reveals how each rule embodies distinctive tradeoffs.

I emphasize that my framework is best viewed as complementary to, not in competition with, the theory of subjective probability. I say nothing about how the agent should take the propositions in which she believes and convert them into choices—the agent could still have a prior over the state space and condition on the propositions in her beliefs to obtain a posterior. When Savage (1954) laid the foundations of subjective expected utility, he was keenly aware of its limitations—he repeatedly acknowledges the infeasibility of acting in accordance with his postulates. In a beautiful passage, Savage describes his preferred solution, contrasting the proverbs "one can cross that bridge when one comes to it" and "look before you leap" to illustrate two extreme approaches to choice. He concludes that

> *[T]o cross one's bridges when one comes to them means to attack relatively simple problems of decision by artificially confining attention to so small a world that the "Look before you leap" principle can be applied there. (pp. 16)*

Savage's advice, in essence, is to limit one's analysis to a "small world" and apply Bayesian reasoning within it. This means articulating a prior over an artificially small state space, which raises new issues.

In this artificially small state space, each state is not a complete description of the world, and the states may not exhaust all possibilities. As a result, an agent who follows Savage's advice has many experiences whose possibility she did not exactly anticipate—she is

[1]See Nickerson (1998) for a survey of evidence on confirmation bias. Within the economics literature, Rabin and Schrag (1999) provide a canonical analysis of confirmation bias. Kahan (2013) and Tappin et al. (2017) offer recent evidence on motivated reasoning.

constantly surprised to varying degrees. Each new experience requires an act of interpretation to fit within her incomplete model. Errors of interpretation, missing details in the description of states, or both, will periodically lead to conflicts. How should a person respond? One approach is to come up with a new model after every surprising observation—imagine a new state space and articulate a new prior every time one is surprised. Another is to augment the state space and extend one's existing prior. Both of these approaches necessarily run into practical limitations of the same sort that lead Savage to "small worlds" in the first place.

I propose a third approach that respects the spirit of small worlds reasoning: try to make sense of the observation within the current model. The finite state space represents the small world our agent has in mind. On hearing a proposition that contradicts her current beliefs, the agent adjusts *within this model*—there is no model revision nor an expansion of the state space. Because of this, beliefs are only tentatively held. We should imagine our agent knows that her model is wrong, and the propositions she "believes" could be false, but without a viable alternative she chooses to act as if they were true. A useful way to reframe my question is: how should a person decide whether she has actually observed an event in her model of the world?

The idea of trust is central. We rarely obtain direct observations of payoff relevant events, so we are forced to rely on what others tell us. When given a new report, one could treat it as a signal, adding new states and forming a prior about the signal's reliability. Alternatively, one could simply trust the report and treat it as a direct observation. We do this routinely in daily life—if my wife tells me it's raining outside, this is as good as seeing rain myself, and I grab an umbrella on my way out.[2] More generally, we typically trust that the data underlying a scientific study is not fabricated, and we trust that our eyes are not playing tricks on us when we see a coin toss come up heads. The point at which we stop second guessing ourselves and accept an observation as an observation can vary across people, but there must come such a point. An update rule here tells the agent how to behave when trusted sources conflict.

Two contributions emerge from the analysis. First, given the severe constraints we face in articulating models of the world to inform our choices, this framework helps us identify tradeoffs inherent in the way we form beliefs. The exercise falls short of defining a "rational" procedure, but ultimately the way we respond to surprises necessarily reflects certain underlying preferences—there is no one right answer. Our preferences may well depend on a broader context, and my framework offers a language in which we can clearly articulate preferences over different tradeoffs.

More speculatively, the paper provides a potentially fruitful approach to modeling boundedly rational agents in applications. In particular, it offers a principled way to represent individuals responding to novel environments, before they acquire large amounts of feedback. The internal logic in positive models of rational choice is that individuals learn from experience what features of the environment are important and how to avoid mistakes, and indeed

[2]Moreover, if I subsequently find sunshine, I do not seriously reconsider my wife's trustworthiness in the future. If this happens repeatedly, the situation changes, but in the absence of more data, I have no reason to view this as anything other than a fluke.

it is standard experimental practice to discard data from several repetitions of a choice problem to ensure that subjects have time to learn. Despite the ubiquity of choices with limited feedback—the three examples with which I began all describe situations in which we rarely see the counterfactual—we thus far lack a theoretical framework in which to analyze such behavior. Much work remains to see if this framework is up to the task, but at the very least it can serve as a tractable building block for new theories.

I offer a new perspective on choice under uncertainty. The data we observe in everyday life are filtered through countless people with ambiguous motivations. Our inability to imagine every possibility and articulate a corresonding prior constrains the way we form beliefs. Some level of trust is necessary if we are to learn anything at all, but then we have to accept the possibility of mistakes—honest or otherwise. My results highlight fundamental tradeoffs. Though there is no obvious "best" way to udpate beliefs, we do have principled options. I proceed immediately with the model exposition and a discussion of the learning axioms. Following the main theorems, I discuss a number of issues and extensions, as well as related work. I conclude with brief remarks.

2 States, Propositions, and Beliefs

There is a finite set of states $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$—assume $n \geq 2$. A subset $P \subset \Omega$, with $P \notin \{\emptyset, \Omega\}$, is a **proposition**. I interpret belief in a proposition as the tentative belief that it contains the true state.[3] If P and P' are such that $P \cap P' = \emptyset$, I say that P and P' **contradict** one another. An agent's **beliefs** are a collection of propositions $B := \{P_i\}_{i \in I}$. If an agent's beliefs are B, then the agent believes that each $P_i \in B$ contains the true state.

An agent learns propositions sequentially. An **information set** is an ordered list of propositions $L = (P_1, P_2, ..., P_k)$. An **update rule** $U(L)$ associates a belief B to each possible list L. Given two lists L and L', let $L + L'$ denote concatenation. I often abuse notation, writing P for a list that contains the single proposition P. Throughout the paper, I impose the following axioms on update rules.

- **Sequential processing:** For any proposition P and any lists L and L', we have $U(L + P) \subseteq U(L) \cup \{P\}$, and if $U(L) = U(L')$, then $U(L + P) = U(L' + P)$
- **Stability:** For any proposition P' and any list L, if there exists $P \in U(L)$ such that $P \subseteq P'$, then $U(L + P') = U(L) \cup \{P'\}$

The first axiom ensures that $U(L)$ describes sequential information processing. Given the list $L = (P_1, P_2, ..., P_k)$, we can imagine arriving at the belief $U(L)$ in k steps. After processing the first ℓ propositions, our agent holds some interim belief B_ℓ—one can view this as what the agent retains in memory after update ℓ. After learning proposition $\ell + 1$, the agent updates to a belief $B_{\ell+1} \subseteq B_\ell \cup P_{\ell+1}$. At each step, the update depends only on

[3]More specifically, this means that in the agent's subjective judgment, she thinks she is better off neglecting the complement of P when making a decision. That is, the agent thinks that contemplating states outside P is not worth the cognitive effort that doing so would entail.

the interim belief B_ℓ, not on the exact list $(P_1, P_2, ..., P_\ell)$ that led to this belief. The second axiom ensures that the notion of beliefs is meaningful: If the agent hears a proposition implied by one that she already believes, she adds this to her beliefs. Note this implies that if told a proposition she already believes, the agent's beliefs do not change.

2.1 Logical Consistency

Coherency requires some restriction on what beliefs the agent can hold. Throughout the paper, I assume there is some set $\mathcal{B}$ of *permissible* beliefs, and $U(L)$ must always take some value in $\mathcal{B}$. There are different restrictions we can consider to define the set of permissible beliefs. The most natural assumption is that each $B = \{P_i\}_{i \in I} \in \mathcal{B}$ is *consistent*, meaning

$$P(B) := \bigcap_{i \in I} P_i \neq \emptyset.$$

If beliefs are consistent, then there exists at least one state in which every proposition the agent believes is true. This suggests a natural equivalence relation between beliefs. I say B and B' are *equivalent*, written $B \cong B'$, if $P(B) = P(B')$.

While I often assume consistent beliefs, it also seems reasonable that a boundedly rational agent might have difficulty detecting inconsistencies. A weaker assumption is k-*consistency*: I say B is k-consistent if there is no collection of $k \geq 2$ propositions $\{P_i\}_{i=1}^k \subset B$ such that

$$\bigcap_{i=1}^{k} P_i = \emptyset.$$

Higher values of k indicate more sophistication. Write $\mathcal{B}_k$ for the set of k-consistent beliefs, and write $\mathcal{B}_\infty$ for the set of consistent beliefs.

2.2 Remarks

The state space represents our agent's model of the world, and this model is coarse. Rather than being a complete description of all that could ever be relevant to the agent, a state simply describes as much detail as the agent can contemplate. This means the agent may face residual uncertainty even when she knows the true state within her model. Alternatively, one might imagine that the agent reasons using a particular language, and this language constrains what she can express.

You should think of propositions as statements about the world that one could express in natural language. For instance, "Ben has a dog" or "the temperature outside is 75 degrees Fahrenheit." Viewed this way, it makes sense to remember particular propositions as opposed to their logical implications. If I later learn that Ben has a cocker spaniel, I do not forget the proposition "Ben has a dog." If I talk to someone about Ben, I may mention that he has a dog without specifying what type. Similarly, if two propositions imply another, I may not infer this without additional prompting. If I know that a lawyer has passed the bar,

and that Ben is a lawyer, it may not immediately occur to me that Ben has passed the bar. Normatively, we might want the agent to make such inferences—one of the learning axioms captures this idea—but the framework can describe agents who are not logically omniscient.

It is important to recognize that a statement from another person need not be understood literally by the agent. Following an interaction with Ben, I might understand the proposition "Ben said that he has a cocker spaniel." The description of a state can include details about who said what, or potentially even how many times the agent has heard something. What is essential is that our agent's capacity to encode such information in her model is *finite*.

The sequential processing axiom embeds two important assumptions. First, all that matters for the update rule is the *set* of propositions our agent believes—whether a given proposition was the first or last that she heard has no effect on updating in the future. Second, after the agent declines to believe a proposition, having previously heard it has no effect on updating in the future. The agent sorts propositions into two classes—those she believes and those she does not—and the axiom precludes differential treatment of propositions within the same class.

Note that some version of this dichotomy between accepting and rejecting a proposition, and a version of the sequential processing axiom, is necessary in any model with bounded memory. If an agent has finitely many memory states and encounters a piece of information, then the agent either transitions to another state or does not, and the way that the agent transitions depends only on the current memory state. I impose additional structure on how the agent uses her finite memory, structure that is consistent with how we typically model uncertainty and information—this is largely inspried by Savage's notion of "small worlds." I discuss ways to interpret this structure in section 5.

2.3 Examples

Here are a few examples of update rules.

The Skeptic Rule

Some people are intensely skeptical of anything they are told. The **skeptic rule** $U_\emptyset(L)$ is identically equal to $\emptyset$ for all lists L. The agent rejects all propositions, maintaining empty beliefs.

Wishful Thinking Rules

Wishful thinking occurs when people believe something because they want it to be true. To capture this idea, consider an ordering $\succ$ on the set of states Ω, and relabel the states so that $\omega_1 \succ \omega_2 \succ ... \succ \omega_n$. Write ω^L for the highest ranked state in $P(U(L))$. The **wishful thinking rule** $U_{WT,\succ}(L)$ updates as follows:

- $U_{WT,\succ}(P) = \{P\}$ for any proposition P
- If P contains a state $\omega \succ \omega^L$, then $U_{WT,\succ}(L + P) = \{P\}$

- If P contains ω^L and no higher ranked state, then $U_{WT,\succ}(L+P) = U_{WT,\succ}(L) \cup \{P\}$
- If $\omega \prec \omega^L$ for all $\omega \in P$, then $U_{WT,\succ}(L+P) = U_{WT,\succ}(L)$

Under a wishful thinking rule, the agent has an underlying preference relation over states and believes whatever propositions contain her most preferred state. If she learns a new proposition containing a more preferred state, she discards every proposition in her beliefs and accepts the new one.

The Stubborn Rule

People often retain their existing beliefs when faced with contradictory evidence. The **stubborn update rule** $U_S(L)$ accepts propositions that are consistent with current beliefs and rejects propositions that conflict with those beliefs. Formally, the stubborn rule satisfies $U_S(L+P) = U_S(L) \cup \{P\}$ if $U_S(L) \cup \{P\} \in \mathcal{B}_\infty$, and $U_S(L+P) = U_S(L)$ otherwise. Unlike the skeptic rule and wishful thinking rules, the stubborn rule is sensitive to the order of propositions in L.

The Blank Slate Rule

Similar to the stubborn rule, the *blank slate rule* $U_B(L)$ accepts propositions that are consistent with current beliefs—we have $U_B(L+P) = U_B(L) \cup \{P\}$ if $U_B(L) \cup \{P\} \in \mathcal{B}_\infty$. However, if $U_B(L) \cup \{P\} \notin \mathcal{B}_\infty$, the blank slate rule satisfies $U_B(L+P) = \emptyset$—the agent discards all propositions and starts over. With this rule, repetition plays a role: the agent believes anything after two consecutive repetitions.

3 Learning Axioms

In this section, I propose several axioms for update rules and argue that they represent desirable properties. For conceptual clarity, I divide the axioms into two categories. The **non-manipulability axioms** capture the idea that a good update rule is invariant across different presentations of the same information. The **willingness-to-learn axioms** capture the idea that a good update rule allows an agent to benefit from hearing information.

I begin with the non-manipulability axioms. For beliefs $B = \{P_i\}_{i \in I}$, recall the notation

$$P(B) = \bigcap_{i \in I} P_i$$

and the corresponding equivalence relation $B \cong B'$ if $P(B) = P(B')$. The non-manipulability axioms are:

- **Order Independence:** If π is a permutation of a list L, then $U(\pi(L)) = U(L)$.
- **Weak Order Independence:** If L and L' are lists, and P and P' are any propositions such that $U(L) \cup \{P\} = U(L') \cup \{P'\}$, then $U(L+P) = U(L'+P')$.

- **Frame Independence:** For any proposition P, if $U(L) \cong U(L')$, then we also have $U(L+P) \cong U(L'+P)$.

Order independence means that beliefs are insentive to the order in which the agent hears propositions—this precludes manipulation by telling the agent the exact same propositions in a different order. Weak order independence applies this principle only within a single update. This axiom requires an agent to treat a proposition the same whether it is new or currently in her beliefs.

Frame independence is more subtle. This axiom says that equivalence is preserved when the agent hears the same proposition following two different lists. If $P(B) = P(B')$, then B and B' are in a sense different ways of framing the same information, and the axiom implies that updating is invariant to this type of reframing. Another way to think about frame independence is as a strong form of logical consistency. Suppose not only that the agent is consistent—the permissible beliefs are $\mathcal{B}_\infty$—but after choosing to believe a proposition, the agent deduces everything implied by that proposition and adds it to her beliefs. Such an agent necessarily follows a frame independent update rule. Likewise, an agent using a frame independent update rule behaves as if she deduces all logical implications of her beliefs.

Turning to the other category, the willingness-to-learn axioms are:

- **Openness:** For any L and any proposition P, there exists L' such that $P \in U(L+L')$.
- **Weak Openness:** For any proposition P, there exists L' such that $P \in U(L')$.
- **Credulity:** For any L and any proposition P, if $U(L) \cup \{P\}$ is a permissible belief, then $U(L+P) = U(L) \cup \{P\}$.

Openness means that, with the right coaxing, our agent can always be convinced of any proposition. Weak openness relaxes this, imposing the condition only when starting from empty beliefs. I find this property desirable largely because of what it rules out. In the absence of openness, the agent may arbitrarily refuse to accept some propositions and may fail to correct mistakes despite ample opportunity.

Credulity means that our agent always believes a proposition if doing so is consistent with what she already believes. If we take a literal interpretation of the model, then this axiom embodies Grice's cooperative principle. The agent presumes that the person giving her information is trying to be helpful, so she errs on the side of trust. There is good reason for this. In the absence of trust, valuable information would go to waste—imagine if every doctor doubted others' medical studies and had to run her own trials before deciding how to treat patients. Alternatively, we could interpret the content of a proposition as exactly what the agent infers from an interaction, in which case credulity becomes indispensable.

Two additional axioms appear in the representation theorem but do not fall neatly into either of the above categories. To state the first, I need some notation. If π is a permutation of the states Ω, and $P = \{\omega_1, \omega_2, ..., \omega_\ell\}$ is a proposition, write $\pi(P) = \{\pi(\omega_1), \pi(\omega_2), ..., \pi(\omega_\ell)\}$ for the proposition obtained by permuting the states in P. For a list of propositions $L = (P_1, P_2, ..., P_k)$, write $L^\pi = (\pi(P_1), \pi(P_2), ..., \pi(P_k))$ for the corresponding list of permuted

propositions. Similarly, for a belief $B = \{P_i\}_{i \in I}$, write $B^\pi = \{\pi(P_i)\}_{i \in I}$ for the set of permuted propositions. The remaining axioms are:

- **Label Neutrality:** For any list L, if π is a permutation of Ω, then $U(L^\pi) = U(L)^\pi$.
- **Conviction:** For any L, any $P \in U(L)$, and any proposition P', if $P \notin U(L + P')$, then $U(L + P') \cup \{P\}$ is not a permissible belief.

Label neutrality means that permuting or relabeling the states has no effect on how the agent processes information. One could interpret this as invariance to the choice of formal language. The axiom could also express a preference for symmetry or a belief that states are equally likely ex-ante. Conviction means that our agent does not discard propositions unnecessarily. If the agent must eliminate something from her beliefs, she makes a minimal adjustment—no throwing the baby out with the bath water. We can view conviction as similar to credulity in that it tries not to waste available information. One could also view conviction as imposing a less tentative meaning for the term "beliefs:" once the agent accepts a proposition, she is committed to it unless forced to reconsider.

4 (Im)possibility Theorems

My first result highlights a tension between order independence and the willingness-to-learn axioms.

Theorem 1 (Impossibility)**.** *If the agent is at least* 2*-consistent—*$\mathcal{B} = \mathcal{B}_k$ *for some* $k \geq 2$*—then*

(a) There is no update rule satisfying both order independence and openness.

(b) If there are at least 5 *states, then there is no update rule satisfying both order independence and credulity.*

Proof. I begin with part (a), assuming for the sake of contradiction that $U(L)$ satisfies openness and order independence. Let ω and ω' be distinct states, and let $\overline{L}$ denote a list containing 2^n copies of all possible propositions—such a list exists because $|\Omega| = n$ is finite. By openness, there exists a list L_ω such that $\{\omega\} \in U(\overline{L} + L_\omega)$, and likewise there exists a list $L_{\omega'}$ such that $\{\omega'\} \in U(\overline{L} + L_{\omega'})$. Consider a permutation of $\overline{L} + L_\omega$ such that all duplicate propositions appear consecutively. I claim that we can remove the propositions in L_ω without changing the resulting beliefs.

To see this, note that when a proposition is repeated multiple times, one of two things can happen. Either at some point the agent accepts the proposition, and stability implies that beliefs cease changing after that, or the agent never accepts the proposition, and $U(L + P) \subseteq U(L) \cup \{P\}$ implies that beliefs eventually cease changing as there are only so many propositions the agent can remove from her beliefs. Since $\overline{L}$ contains 2^n copies of every proposition, we are sure to arrive at one of these outcomes before processing repetitions

due to the propositions in L_ω. I conclude by order independence that $U(\overline{L} + L_\omega) = U(\overline{L})$. Similarly, we have $U(\overline{L} + L_{\omega'}) = U(\overline{L})$. This means that both $\{\omega\} \in U(\overline{L})$ and $\{\omega'\} \in U(\overline{L})$, which contradicts 2-consistency.

For part (b), suppose again for the sake of contradiction that $U(L)$ satisfies credulity and order independence. Suppose $\omega_1, \omega_2, \omega_3, \omega_4, \omega_5 \in \Omega$ are distinct, and define $P_i = \{\omega_i, \omega_{i+1}\}$ for $i = 1, 2, 3, 4$ and $P_5 = \{\omega_5, \omega_1\}$. Given P and P' disjoint, write $P \succ P'$ if $U(P + P') = U(P' + P) = \{P\}$. The proof is based on the following claim: if P_i and P_j overlap, and P_k is disjoint from both then either both $P_i \succ P_k$ and $P_j \succ P_k$ or both $P_k \succ P_i$ and $P_k \succ P_j$.

Assuming the claim, we can quickly construct contradiction. Suppose $P_1 \succ P_3$. Applying the claim repeatedly gives

$$\begin{aligned} P_1 \succ P_3 \implies P_1 \succ P_4 \implies P_2 \succ P_4 \\ \implies P_2 \succ P_5 \implies P_3 \succ P_5 \implies P_3 \succ P_1, \end{aligned}$$

a contradiction. I conclude that there is no update rule satisfying both order independence and credulity.

To establish the claim, I first show that we cannot have $U(P_i + P_k) = \emptyset$. If this were true, then credulity and order independence imply

$$U(P_i + P_k + P_j) = \{P_j\} = U(P_j + P_k + P_i).$$

Credulity and sequential processing imply that if the right hand side contains P_j, it must also contain P_i, a contradiction. Similarly, we cannot have $U(P_j + P_k) = \emptyset$.

Now suppose that $U(P_i + P_k) = \{P_k\}$ but $U(P_j + P_k) = \{P_j\}$. Credulity and sequential processing imply that

$$U(P_i + P_k + P_j) = U(P_k + P_j) = \{P_j\}, \quad \text{and} U(P_j + P_k + P_i) = U(P_j + P_i) = \{P_j, P_i\},$$

contradicting order independence. The claim follows. □

Thinking first about part (a), order independent update rules should intuitively resemble the wishful thinking rule. If the agent is not a skeptic, then $U(\{\omega\} + \{\omega'\}) = U(\{\omega'\} + \{\omega\})$ implicitly defines an ordering on states that is incompatible with openness. Part (b) relies on having 5 states. Figure 1 offers a visual depiction of the key step. I essentially construct an ordering on propositions that contradict one another—write $P_i \succ P_j$ if $U(P_i + P_j) = U(P_j + P_i) = P_i$. The proof rests on the observation that if proposition P_k is disjoint from both P_i and P_j, then order independence implies either $P_k \succ P_i$ and $P_k \succ P_j$, or $P_k \prec P_i$ and $P_k \prec P_j$. Applying this rule within a cycle of overlapping propositions yields a contradiction.

There are two especially striking features of this result. First, we have impossibility for key *pairs* of axioms. Forget satisfying every desirable property for an update rule—we cannot even achieve any pair of axioms. This points to a serious dilemma between facilitating learning and avoiding manipulation. Second, this finding relies only on the weak assumption of 2-consistency. Even if the agent can entertain a wide range of inconsistent beliefs, as long as she never believes two propositions that directly contradict one another, the conclusion of Theorem 1 holds.

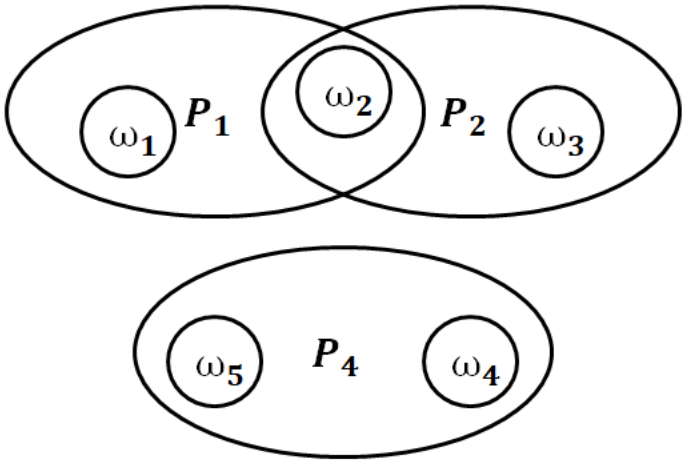

Figure 1: We either have both $P_1 \succ P_4$ and $P_2 \succ P_4$, or both $P_4 \succ P_1$ and $P_4 \succ P_2$.

Despite this negative result, there are many combinations of non-manipulability and willingness-to-learn axioms that update rules *can* satisfy. Section 2.3 presented several examples. The skeptic rule satisfies all three non-manipulability axioms. A wishful thinking rule is order and frame independent, and it satisfies weak openness. The stubborn rule satisfies label neutrality, frame independence, credulity, and conviction. As a counterpoint to the impossibility theorem, my second result traces the boundary of what is possible, highlighting specific update rules and reasons to adopt them.

To state the result, I need to introduce a variation on the stubborn rule. A **nearly stubborn** update rule U has the following structure:

- If $P \cap P(U(L)) \neq \emptyset$, then $U(L + P) = U(L) \cup \{P\}$
- If $P \cap P(U(L)) = \emptyset$ and $|P| \geq 2$, then $U(L + P) = U(L)$
- If $\{\omega\} \cap P(U(L)) = \emptyset$ and $|P(U(L))| = 1$, then $U(L + \{\omega\})$ contains $\{\omega\}$ and every $P \in U(L)$ such that $\omega \in P$.
- If $\{\omega\} \cap P(U(L)) = \emptyset$ and $|P(U(L))| = k > 1$, then either $U(L + \{\omega\})$ contains $\{\omega\}$ and every $P \in U(L)$ such that $\omega \in P$, for any state ω and any L such that $|P(U(L))| = k$ and $\{\omega\} \cap P(U(L)) = \emptyset$, or $U(L + \{\omega\}) = U(L)$ for any state ω and any L such that $|P(U(L))| = k$ and $\{\omega\} \cap P(U(L)) = \emptyset$.

The first bullet is exactly the credulity axiom, and the second says that nearly stubborn rules behave like the stubborn rule whenever a new proposition contains at least two states. Differences emerge only if the new proposition is a singleton, and the third and fourth bullets cover this case. If the current beliefs pin down a single state, then a nearly stubborn rule must accept a new proposition if it is a singleton. If the current beliefs do not pin down a single state, a nearly stubborn rule may or may not accept a singleton that contradicts the

current beliefs, but whether it accepts the new proposition or not can only depend on the number of states that are consistent with current beliefs, not on state labels.

The last two bullets may appear somewhat strange at first glance, but on closer inspection, they are quite natural. The third bullet ensures that nearly stubborn rules satisfy openness, and by avoiding updates that depend on state labels, the last two bullets also ensure that nearly stubborn rules satisfy label neutrality. Note there is more than one such update rule—for instance, the rule that only accepts singleton propositions if current beliefs pin down a single state is one nearly stubborn rule, but another such rule always accepts singleton propositions.

Theorem 2 (Possibility)**.** *Suppose the agent is consistent—*$\mathcal{B} = \mathcal{B}_\infty$.

(a) The update rule U satisfies label neutrality and order independence if and only if U is the skeptic rule.

(b) The update rule U satisfies order independence, frame independence, and weak openness if and only if U is a wishful thinking rule for some order on Ω.

(c) The update rule U satisfies label neutrality, weak order independence, frame independence, openness, and credulity if and only if U is the blank slate rule.

(d) The update rule U satisfies label neutrality, frame independence, openness, credulity, and conviction if and only if U is a nearly stubborn rule.

Proof. For part (a), it should be clear that the skeptic rule satisfies label neutrality and order independence. Going the other direction, suppose there exists a proposition P such that $U(P) = \{P\}$. By label neutrality, we have $U(P') = \{P'\}$ for any P' such that $|P'| = |P|$. Let L denote a list containing all such P'. Label neutrality and order independence imply that for any permutation π of Ω, we have

$$U^\pi(L) = U(L^\pi) = U(L).$$

This implies that $U(L) = \emptyset$—if there exists $\omega \in P\left(U(L)\right)$ and $\omega' \notin P\left(U(L)\right)$, then choose π that swaps ω and ω' to get a contradiction. Sequential processing implies that $U(L+P) = P$, but order independence and stability imply that $U(L+P) = \emptyset$—just move the two copies of P to the front of the list—a contradiction. Hence, the skeptic rule is the only rule satisfying label neutrality and order independence.

For part (b), I first note that the definition of a wishful thinking rule immediately implies that it satisfies order independence, frame independence, and weak openness—the beliefs $U(L)$ consist of every proposition that contains the highest ranked state that appears in some proposition in L. Now suppose an update rule satisfies order independence, frame independence, and weak openness. I first claim that $U(P) = \{P\}$ for any proposition P. Since U satisfies weak openness, there exists a list L such that $P \in U(L)$. By order independence, we end up with the same beliefs if we reorder L so that all copies of P appear at the start. Hence, we must have $P \in U(P)$.

Now, consider a list L^* containing every singleton proposition. I claim that $U(L^*)$ is non-empty. If L^* were empty, then by the previous claim and sequential processing, we have $U(L^* + P) = \{P\}$ for any singleton proposition. Since we can reorder $L^* + P$ so that both copies of P appear at the start, order independence and stability now imply that $U(L^*) = P$, a contradiction. Hence, we have $U(L^*) = \{\{\omega_1\}\}$ for some state ω_1, and order independence implies that U returns this belief from any list of singletons containing $P = \{\omega_1\}$. The state ω_1 is our candidate for the highest ranked state in our ordering on Ω.

I next claim that if there exists $P \in L$ with $\omega_1 \in P$, then $U(L)$ contains all propositions in L that contain ω_1. I proceed in three steps. First, I show $\{\omega_1\} \in U(P + \{\omega_1\})$ for any proposition P. Suppose $\omega_1 \in P$. Order independence and stability imply that

$$U\left(P + \{\omega_1\}\right) = U\left(\{\omega_1\} + P\right) = \{\{\omega_1\}, P\}.$$

Now suppose $\omega_1 \notin P$, and consider some other state $\omega \in P$. We have

$$U\left(P + \{\omega_1\} + \{\omega\}\right) = U\left(\{\omega\} + P + \{\omega_1\}\right) \cong U\left(\{\omega\} + \{\omega_1\}\right) = \{\{\omega_1\}\},$$

where the first equality follows from order independence, and equivalence follows from frame independence and stability. Since $U\left(P + \{\omega_1\} + \{\omega\}\right) \cong \{\{\omega_1\}\}$, sequential processing implies $\omega_1 \in U\left(P + \{\omega_1\}\right)$.

The second step shows that $\{\omega_1\} \in U(L)$ whenever $\{\omega_1\} \in L$. Due to order independence, this is equivalent to showing that $\{\omega_1\} \in U(L + \{\omega_1\})$ for any list L. Frame independence implies that

$$U\left(L + \{\omega_1\}\right) \cong U\left(P(U(L)) + \{\omega_1\}\right),$$

and the previous step implies that $\{\omega_1\}$ is contained in the right hand side. If $\omega_1 \notin P(U(L))$, we are done—beliefs must contain $\{\omega_1\}$. If $\omega_1 \in P(U(L))$, then we can divide the propositions in L into two lists L_0 and L_1, such that L_0 contains all propositions that do not contain ω_1, and L_1 contains all propositions that contain ω_1. By order independence, we have

$$U\left(L + \{\omega_1\}\right) = U\left(L_0 + \{\omega_1\} + L_1\right).$$

We know that $\{\omega_1\} \in U\left(L_0 + \{\omega_1\}\right)$ from our earlier work, so stability now implies the right hand side contains $\{\omega_1\}$.

The third step shows that if $\omega_1 \in P$, then $P \in U(L)$ for any list L containing P. By order independence, this is equivalent to showing that $P \in U(L + P)$ for any L. Observe that $\{\omega_1\} \in U\left(L + \{\omega_1\} + P\right)$, so stability implies that $P \in U\left(L + \{\omega_1\} + P\right)$, and order independence now implies $P \in U\left(L + P + \{\omega_1\}\right)$. Sequential processing now implies $P \in U\left(L + P\right)$ as desired.

To show that U treats ω_1 as the highest ranked state of a wishful thinking rule, we need that $P \notin U(L)$ whenever $\omega_1 \notin P$ and $\omega_1 \in P'$ for some $P' \in L$. By order independence, it is enough to show that $P \notin U(P + P')$ whenever $\omega_1 \notin P$ and $\omega_1 \in P'$. If $P \cap P' = \emptyset$, our work above already implies the claim, so assume $P \cap P' = P^* \neq \emptyset$. Choose $\omega \notin P'$ distinct from ω_1. Frame independence implies

$$U(P + P' + \{\omega_1, \omega\}) \cong U(P^* + \{\omega_1, \omega\}) = \{\omega_1, \omega\},$$

but our earlier work implies

$$U(P + P' + \{\omega_1, \omega\}) = U(P' + \{\omega_1, \omega\} + P) \cong U(\{\omega_1\} + P) = \{\omega_1\},$$

a contradiction. I conclude that U treats ω_1 as the highest ranked state of a wishful thinking rule.

To complete the proof of part (b), we induct on the states. Repeat the above argument starting with a list L^* that contains all singletons *except* $\{\omega_1\}$, and considering only propositions that do not contain $\{\omega_1\}$. This singles out a state $\{\omega_2\}$ that the update rule treats as the second highest ranked state. Repeating until all states are exhausted finishes the argument.

For part (c), it should be clear the blank slate rule satisfies the stated properties. Suppose U is a rule satisfying label neutrality, weak order independence, frame independence, openness, and credulity. Credulity implies the agent accepts all propositions that are consistent with current beliefs. Hence, we need only show that $U(L+P) = \emptyset$ whenever $P(U(L)) \cap P = \emptyset$. I proceed in three steps. First, it follows from label neutrality and weak order independence that $U(P+P') = \emptyset$ whenever $P \cap P' = \emptyset$ and $|P| = |P'|$—otherwise we obtain a contradiction by choosing a permutation that swaps each state in P with one in P'.

Step two shows that $U(P+P') = \emptyset$ whenever $P \cap P' = \emptyset$. By weak order independence, it is without loss of generality to assume $|P| < |P'|$. Choose $P^* \supset P$ such that $|P^* \cap P'| = |P|$, and write $P_* = P^* \cap P'$. Credulity and frame independence imply $U(P+P') \cong U(P^*+P+P')$. Credulity and weak order independence imply that $U(P^* + P + P') = U(P^* + P' + P)$. Credulity and frame independence imply $U(P^* + P' + P) \cong U(P_* + P)$, and step one implies that $U(P_* + P) = \emptyset$. Finally, to show that $U(L + P) = \emptyset$, observe that frame independence implies $U(L + P) \cong U\left(P(U(L)) + P\right) = \emptyset$.

For part (d), I first verify that nearly stubborn rules satisfy the given axioms. Frame independence, credulity, conviction, and label neutrality are immediate from the definition. To see that they satisfy openness, choose a state $\omega' \in P(U(L))$, any P, and a state $\omega \in P$—we have $P \in U(L + \{\omega'\} + \{\omega\} + P)$ for any L. To go the other direction, first note that credulity implies the agent accepts any proposition that is consistent with her existing beliefs—the first bullet in the definition is immediate.

I next show that if $|P| \geq 2$ and $P(U(L)) \cap P = \emptyset$, then $U(L + P) = U(L)$. Suppose there exist L and P that violate this claim. If $U(L + P) \neq U(L)$, then since $U(L) \cup \{P\}$ is not permissible, there must exist $P' \in U(L)$ such that $P' \notin U(L + P)$. Conviction then implies that $P \in U(L + P)$ because otherwise $U(L + P) \cup \{P'\}$ is permissible. By frame independence, we now have

$$U(L + P) \cong U\left(P(U(L)) + P\right) = \{P\}.$$

Take distinct states $\omega, \omega' \in P$, and consider the proposition $P' = P(U(L)) \cup \{\omega\}$. By credulity, we have $U\left(P(U(L)) + P'\right) = \{P(U(L)), P'\}$. Credulity and conviction imply that either

$$U\left(P(U(L)) + P' + P\right) = \{P(U(L)), P'\} \cong \{P(U(L))\}, \text{ or}$$
$$U\left(P(U(L)) + P' + P\right) = \{P', P\} \cong \{\omega\}.$$

Frame independence implies $U\left(P(U(L)) + P' + P\right) \cong U\left(P(U(L)) + P\right) = P$, a contradiction. Therefore, the agent must reject P when it appears after L, and conviction implies the agent cannot discard any proposition in $U(L)$. The third bullet follows immediately from openness, label neutrality, and frame independence, and the last is immediate from label neutrality and frame independence.

□

Different axioms may hold more or less appeal depending on context, and Theorem 2 highlights what rules follow from different tradeoffs. Parts (a) and (b) highlight the implications of the non-manipulability axioms. Order independence together wtih a symmetry condition precludes any form of learning. Wishful thinking appears as a natural response to concerns about manipulation. Here, the weak openness axiom pins down a unique rule, but it should be clear that the wishful thinking rule is not the only non-trivial update rule that satisfies both order and frame independence. For instance, suppose we adjust a wishful thinking rule so that there is some set of propositions S that are never added to beliefs—to obtain the beliefs $U(L)$, we delete from L any proposition contained in S to obtain a list L', and apply the wishful thinking rule to the list L'. Any such rule satisfies order and frame independence, and an argument similar to that in part (b) shows that any update rule satisfying order and frame independence must take this form.

Parts (c) and (d) call attention to the willingness-to-learn axioms. Both the blank slate rule and the nearly stubborn rules satisfy openness and credulity. Weak order independence is the key property that leads to the blank slate rule—if the agent can favor neither current beliefs nor the new proposition when facing a contradiction, then the only option is to discard everything—while conviction is the essential property behind nearly stubborn rules. More generally, the two willingness-to-learn axioms permit a large range of different update rules. Although credulity dictates what happens when a new proposition is consistent with current beliefs, the update rule can do almost anything when facing a contradiction—openness imposes only a minimal restriction, analogous to the third bullet in the definition of nearly stubborn rules, that the update rule accept *some* contradictory propositions when current beliefs are equivalent to a single state.

Unlike Theorem 1, this characterization of update rules clearly depends on the agent having consistent beliefs. Moreover, frame independence implies a stronger form of logical consistency, and we needed this to pin down the rules for parts (b), (c), and (d). Part (a) is an exception, but relaxing consistency changes little of substance here. Suppose the agent were 2-consistent. To satisfy label neutrality and order independence, the agent could accept all propositions containing at least $\lfloor \frac{n}{2} \rfloor + 1$ states and reject all others—any two such propositions must overlap. In this case, the lack of logical consistency raises questions, beyond the scope of this paper, about how the agent might map beliefs to actions.

5 Discussion

This section discusses in turn ways to interpret the model, how the model relates to choice rules, how I view the contributions of this paper, and several directions to extend the analysis.

5.1 Interpreting the Model

I first highlight a literal interpretation in which we view propositions as direct statements, and our agent chooses whether to believe the literal meaning of those statements. I subsequently discuss a personalistic interpretation in which we view a proposition as the subjective inference an agent draws from interacting with an information source—based on a combination of context, language, and non-verbal cues, a given interaction may allow the agent to rule out certain states within her model. In both cases, we view the notion of a belief as a tentatively held attitude. Because the agent's model is incomplete, she is necessarily prepared to accept that something she currently believes turns out to be false. This is consistent with the epistemological stance of critical rationalism—all belief is conjecture, belief is never justified, and certainty is impossible (Popper, 1963). The main problem our agent faces is precisely how to revise her beliefs when faced with contradictions.

A Literal Interpretation

Here is a thought experiment. Imagine you are seated at a table, and across from you a person flips five coins and hides them behind a cover. Two envelopes are pushed towards you. The left envelope contains a dollar for every heads, and the right contains a dollar for every tails. You get to choose one to take. While you do not see the coins, a sequence of strangers approaches the other side of the table. One by one, each observes the coins and makes a statement about them—e.g. "the first two coins are heads." If one just models the five coins, there are 32 states in Ω, corresponding to every possible ordering. The strangers' statements are propositions that explicitly exclude some states from the realm of possibility. Absent incentives to mislead, one might reasonably trust what the strangers say—if told that the first two coins are heads, one could think that the left envelope contains three and a half dollars in expectation. This seems fine if each new statement is consistent with all prior ones, but what if some statements contradict one another?

The conventional response is to imagine a larger universe of states. A state must now include a description of reasons why the strangers across the table might either lie or misperceive the coins' status. You need to assign prior probabilities to each of these reasons. However, without some knowledge of the strangers across from you, such an exercise is necessarily arbitrary and speculative. It is also cognitively costly.

According to my framework, you choose not to engage in speculation. Instead, you pick some (or none) of the statements to trust, make your choice, and move on with your day. How should we think about sequential processing here? Maybe you do not take care to count repetitions, or you do not recall the order in which strangers made their state-

ments. The assumption is stark, but it seems reasonable to expect a coarse assessment of the statements—nothing substantive changes if you can remember whether a given statement was made more or less than five times, for instance. This interpretation makes the most sense when there is little we can infer from context, and statements have a clear literal meaning. Blind speculation about why such statements conflict seems unproductive.

A Personalistic Interpretation

Given context, a person may understand a statement as something besides or beyond its literal meaning. In the personalistic interpretation, we view a proposition as what results from the agent's subjective understanding of an interaction. The agent translates her experience into the language of her model, and the content of an article, a conversation, or a tweet, is then defined by the set of states it allows the agent to rule out.

There are at least two reasons why a person's understanding of an interaction may differ from the exact words that are exchanged. One is that we make certain inferences based on the norms surrounding conversation. The philosopher Paul Grice called attention to these norms through his elaboration of the "cooperative principle." Grice's insight was that listeners make certain presumptions of a speaker in typical settings. We presume a speaker is being truthful (maxim of quality), is providing all needed information, and no extra (maxim of quantity), is being relevant to the topic at hand (maxim of relation), and is trying to be clear (maxim of manner). Speakers in turn exploit these presumptions to convey meaning beyond their literal words. As one example, consider the exchange

A: I'm out of gas.

B: There's a gas station around the corner.

Typically, A would conclude that the gas station around the corner is open because A presumes that B is trying to help.

Because of these norms, repetition of a statement by the same person may not entail repetition of a proposition as understood by the agent. Suppose I run into a colleague in the morning who tells me "it's supposed to rain later today." I might understand this as meaning "there is a greater than 50 percent chance it will be raining when I leave the office." Now suppose I subsequently see this colleague after lunch, and the colleague repeats to me "it's supposed to rain later today." From this I might understand that my colleague has obtained more information in the interim, and the repetition is intended to convey that the chance of rain is much higher. Alternatively, I might understand this as an indication of importance—not only is it going to rain, but it will rain hard enough that I need an umbrella.

In addition to these inferences, a person may also distinguish different sources. If a friend excitedly tells me about a nutritional supplement with anti-aging properties, I may not understand this as evidence of anti-aging properties, but I might rule out states of the world in which the supplement is toxic. Likewise, my understanding of a statement may vary with the motives of the speaker. The word "quaint" means different things coming from a realtor versus a friend.

One might complain this makes the personalistic interpretation too flexible and precludes meaningful conflict between propositions as understood by an agent, but this is far from the case. First, even information provided sincerely may contain mistakes. A proof in a paper might contain an error. A witness to a murder might misremember the timeline. Second, sources that we generally consider trustworthy often issue conflicting advice. Regulators in the United States and Europe differ in their assessments about whether some ingredients are safe for human consumption. Doctors disagree about whether it is safe to consume sushi during pregnancy. Many interactions convey something surprising or unexpected about the world that makes it difficult to decide just what to believe afterwards.

5.2 Beliefs and Choice

An important feature of the framework is that beliefs, not preferences, are primitive. When I say that the agent "believes" a proposition, I mean that the agent has decided to neglect its complement when making choices. As a result, the model of belief formation is orthogonal to how the agent translates beliefs into choices. One obvious possibility is that the agent has a prior over Ω, and given beliefs B, she conditions on $P(B)$ to obtain a posterior and makes choices to maximize expected utility with respect to this posterior. The beliefs B say nothing about what probability to assign different states. Rather, they tell the agent what events she has observed—the support of the prior—and we can understand the update rule as a way to address zero probability events.

While pairing this framework with expected utility maximization is natural, it is by no means necessary. As one alternative, consider minimax regret. The minimax regret criterion is "prior free," but it is highly sensitive to the states over which regret is assessed—an imaginative agent could envision a state to justify any choice in a particular situation. Paired with this decision rule, the beliefs B would constrain the states over which regret is computed—belief in a proposition means believing it is safe to exclude its complement when evaluating an action according to the minimax criterion. We could just as easily pair the framework of this paper with other choice rules like minimax expected utility with multiple priors or the Hurwicz criterion.

5.3 Towards a Normative Theory of Bounded Rationality

One way to understand the framework in this paper is as a step towards a normative theory of rationality that respects cognitive limitations. A normative theory of choice and belief formation needs to tell a user both how to respond to information whose possibility was anticipated and how to respond to information whose possibility was not anticipated. If we assume the agent imagines a state space to represent the world, then responding to anticipated information means evaluating choices after observing an element in some partition of the state space, and decision theory offers an abundance of appealing ways to do this (e.g., expected utility maximization, maximin expected utility over multiple priors,

minimax regret).[4] Cognitive limitations necessitate dealing with the unexpeted. When choosing how to respond to unanticipated information, we face a trilemma: i) we can *change paradigms*, imagining an entirely new model of the world, ii) we can *accommodate* the new observation by extending our existing model with additional states, or iii) we can proceed with *business-as-usual*, interpreting the surprise within our existing model. I follow the third horn in this paper.

Existing work on responding to surprises adopts either the first or second horn, but there are at least two reasons why we need an approach that follows the third horn of the trilemma. The first stems from practical considerations: coming up with a new model in response to every surprise is infeasible. If a person must proceed with business-as-usual at least *some* of the time, then a complete normative theory needs to offer guidance on what to do in these instances. The second is that, on purely normative grounds, coming up with a new model is not always appealing.

Suppose you are deciding whether to wear a face mask in public during a pandemic, and your initial model contains two states: masks help prevent transmission or not. Imagine that in February, the CDC tells you that face masks do not prevent transmission, but later in April the CDC reverses this guidance.[5] What should you believe? One response is to expand the state space in your model—perhaps the CDC had weak evidence for face masks in February but strong evidence in April, so you add states to describe the quality of evidence and the time at which it became available. Alternatively, you might revise your model more radically and speculate that the CDC lied in February in an effort to protect supplies for healthcare workers.[6] Your new model would include states describing various ulterior motives along with whether face masks are effective at preventing transmission. Both of these stories suggest accepting the new guidance, that face masks are effective, but one could just as easily have imagined a new model to support the original guidance—perhaps the new evidence is faulty, or the reversal is due to political pressure on the CDC to look like they are taking action. Since we can tell ourselves a story to justify either choice—wearing a facemask or not—this suggests the question: why bother articulating a new model in the first place?

Coming up with a new model might make sense if there were a naturally given state space, but here one's imagination is the only thing limiting the possible states that are reasonable to consider. In statistical models, assumptions of independence or exchangability greatly simplify the task of articulating a prior, but here there is no reason to impose such assumptions on the sequence of press releases, so articulating a prior entails a great deal of work. If the only purpose of this model is to decide whether to wear a mask or not, why not cut out the middle step and save the effort?[7] I am not arguing that we should never revise or extend our current model of the world, far from it, but there are instances in which doing

[4]Stoye (2011) provides a detailed accounting of different procedures, together with their axiomatic foundations.

[5]This happened during the COVID-19 pandemic.

[6]At present, it appears that this is in fact the reason for the earlier guidance.

[7]This is not to suggest that uncovering the actual reasons for the guidance reversal is not important to do, but the typical person making a choice about whether to wear a mask on a daily basis does not need to be the one doing this.

so makes little sense. Again, if we are to proceed with business-as-usual at least *some* of the time, we should have a normative theory for what to do.

This paper does not provide a definitive answer, but it offers a framework that can help with preference articulation—I highlight the inherent tensions between different desirable properties and the implications of various tradeoffs one might make. In my view, this is as much as one can expect from any theory of rationality. As Gilboa et al. (2009) argue, different ideas of what rational behavior should entail are often in conflict, and "the question is not "what is *the* rational thing to do?" but "what is more rational to do in this instance?"" Different principles may hold more or less appeal in different situations, but to make thoughtful choices, we need a language in which to express our options.[8]

5.4 Rational Behavior in the Absence of Feedback

When we assume rational behavior in descriptive models, we implicitly presume that people in the relevant situations have had ample opportunity to learn from feedback and experience in similar settings. For the models to make sense, people need to be able to learn not only how to interpret information sources but also what features of the environment are important—they must learn a model. However, even without the benefit of experience in similar situations, most information we obtain is delivered in familiar language with a clear meaning, and this meaning serves as a focal point for how we think about a problem. We might not know about reasons why the information could be erroneous or incomplete, but we can always choose to either take the information at face value or ignore it.

A different way to understand the framework in this paper is as a tool to model how people respond to information when feedback and experience are limited. The paper does not take a stance on what the right model is—without any restrictions, the update rules can capture a wide range of behavior—but the axiomatic results can offer practical guidance for where to start. When data are unavailable to discriminate between positive theories—either because theories have not yet been formulated and tested, all available theories have known flaws, or key parts of a theory cannot be tested in practice—axiomatizations can serve as a useful tool to make a considered choice (Gilboa et al., 2019).

5.5 Further Directions

As in all models, several potentially important features are missing. Repetition of statements seems important for how people form beliefs in practice, yet the model seems to preclude a role for repetition. This is less of an issue in the personalistic interpretation because the agent may understand subsequent repetitions as meaning different things. Taking the literal interpretation, while people *do* respond to repetition, it is not clear that they *should*. Nevertheless, the blank slate rule provides one example in which repetition can have an effect, and its axiomatization reveals how this feature derives from more basic

[8]Gilboa et al. (2009) are principally concerned with rational *choice*, whereas I am principally concerned with rational *beliefs*, but the reasoning behind their argument is similar in spirit to mine.

properties—credulity and weak order independence—that we might want an update rule to satisfy.

The limited role for repetition ultimately stems from the restriction to two levels of belief in a proposition. The agent can accept a proposition or not. There is no way to have a stronger or weaker belief in it. Suppose instead that we allow the agent three levels of belief. Imagine there are propositions of which the agent is reasonably sure (category 1), those she thinks likely (category 2), and everything else. In such a model, repetition might become an important part of how propositions move from category 2 to category 1. While not immediately clear how to adapt some of my axioms to this richer setting (e.g. credulity), it is still evident that order independence and openness are in conflict, and it is not hard to generalize my characterization of skepticism—an outline of this extension is available on my personal website. The fundamental issues are still present.

In the literal interpretation, it seems important to account for different levels of trust in different sources. An appealing extension would attach a source to each proposition, allowing the agent to treat the same proposition differently depending on who said it. If the number of distinguishable sources is finite, this is in one sense equivalent to enlarging the state space. However, the added structure may suggest additional axioms that capture something important.

6 Related Work

My framework relates mainly to two strands of literature—work on responding to surprises and work on bounded rationality. On responding to surprises, authors typically take one of two approaches. The first is model revision. In this approach, an agent appeals to some meta-model if her current model proves unsatisfactory. For instance, Ortoleva (2012) axiomatizes a hypothesis testing procedure in which an agent is Bayesian, but after observing an event that is sufficiently unlikely given her model, she considers a collection of priors, selecting the prior that assigned the highest likelihood to the unexpected observation. Relatedly, Galperti (2019) studies how a sender might leverage these "changes in worldview" for persuasion. In an alternative approach, Radner (2002) studies how to estimate a kth order Markov process when k is unknown. He finds a consistent estimator that involves estimating the parameters for a given k and periodically increasing the value of k under consideration. A second approach is to expand the underlying state space to accommodate new possibilities. As one example, Karni and Viero (2013) propose axioms that lead to "reverse Bayesianism." In their model, an agent can become aware of new states or acts and extends her prior in a way that preserves likelihood ratios in the old state space.

My model attempts to address shortcomings of this literature. If we consider the idea that an agent should revise or expand her model in response to a surprise, we must presume that either a different model is readily available, or the agent is prepared to articulate one on the spot. Viewed normatively, this runs into the issues I discuss in Section 5.3. Viewed descriptively, this runs a risk of building vacuous theories without some restriction on the space of permissible models. Though admittedly simplistic, the alternative approach in this

paper leads to update rules that are clearly feasible for cognitively constrained agents, and theories that build on this framework should be falsifiable.

This paper fits into a broader context of research that studies decision making when agents face cognitive limitations or have imperfect models of the environment. Within this literature, Wilson (2014) is closest in spirit to my work, representing an agent as a finite automaton that is optimally adapted to the environment given the finite memory constraint. This approach implicitly assumes that the agent has access to a great deal of experience and feedback in similar choice problems. In contrast, my agent has a preconceived model of the world, and I propose interpretable axioms on how the agent uses her model—I view my approach as a useful starting point to think about reasoning with finite memory in the absence of rich experience and feedback.

Rubinstein and Salant (2006) provide inspiration for the procedural element in my model. They study choice among alternatives presented as an ordered list, and natural axioms lead to order dependent choice rules. I adopt a similar approach to belief updating. An agent processes information as an ordered list of propositions, and the list order often affects beliefs. However, belief updating is different from choice among alternatives in important ways—an agent may believe multiple propositions simultaneously, and there are consistency constraints on which ones. My results complement the earlier paper, enhancing our understanding of how procedural aspects of choice and belief formation affect decision making.

Within the literature on bounded rationality, relatively few papers directly address the failure of logical omniscience. Lipman (1999) provides an important exception. Lipman introduces a framework in which propositions are primitive, and a state is a collection of propositions that completely describe the world. The framework endogenizes a subjective state space, allowing an agent to entertain states comprising inconsistent collections of propositions. Lipman writes in the introduction that: "The standard interpretation of a state of the world is as a complete, consistent description of a way the world might be. The logical omniscience assumption is precisely the consistency part of this definition." My perspective differs in that I also view the completeness part as problematic. Consequently, in my framework I explicitly interpret states as incomplete descriptions of the world.

Beyond the economics literature, my framework vaguely resembles the AGM theory of belief revision in propositional logic (Alchourrón et al., 1985).[9] This theory represents beliefs as the logical closure of a set of propositions and axiomatizes a rule in which one always accepts a new proposition, adding it to the existing knowledge base if it is consistent and otherwise discarding the old knowledge base—this prescription resembles my blank slate rule, except the blank slate rule does not retain the newest proposition. My approach differs in that I do not presume that the new proposition should be seen as more reliable than the old knowledge base, and I propose an entirely different collection of axioms based on normative, rather than aesthetic, considerations.

I view the present paper as a companion to Sadler (2020), which studies opinion and

[9]Basu (2019) uses the AGM theory as a starting point to consider updating in response to zero probability events. The earlier paper focuses on how to assign probabilities to states, not how to determine the support of the posterior, and it does not consider the possibility of contradictory observations.

belief influence in large networks of agents. The spread of false beliefs is intimately tied to the structure of social networks, and the other paper explores how such structures determine the influence agents exert on others' opinions. One can view the model in the other paper as an application of the framework I introduce in this paper—in essence, agents in a network have random opportunities to communicate propositions to their neighbors. The other paper highlights how this model can offer practical guidance in the design of influence campaigns.

7 Final Remarks

How do, and how should, people decide what is true? Economics typically sidesteps this question, assuming that the information content of an observation is self-evident. However, the current political and social media landscape, and the increasing salience of false beliefs, demands that we address situations in which it is not. I offer a formal framework to study these questions and provide a step towards understanding the practical limitations of what we can do.

The necessity of taking some statements on faith together with the existence of falsehoods highlights new issues in belief updating. There is an inherent tension between axioms that prevent manipulation and those that allow learning. In practice, we do trust at least some of what people tell us, and the results of Section 4 highlight the vulnerabilities this can create. Moreover, the representation theorem uncovers deep connections between key principles and common heuristics—wishful thinking is a natural response to the threat of manipulation, and stubbornness is a natural consequence of strong trust in others.

I believe there are many promising directions for future work. In my view, applications to communication games, particularly involving players with different levels of sophistication, seem especially intriguing. These applications may improve our understanding of phenomena like belief polarization, echo chambers, and search for confirming information. Beyond this, I believe my model can offer a useful building block in a richer analysis of collective belief formation. People learn through their experiences to trust different sources, and this suggests that differences in the kind of information that individuals *can* verify on their own, and differences in social connections, should play a key role in determining what becomes "true" within different communities.

References

Alchourrón, Carlos, Peter Gärdenfors, and David Makinson (1985), "On the Logic of Theory Change: Partial Meet Contraction and Revision Functions." *The Journal of Symbolic Logic*, 50, 510–530.

Basu, Pathikrit (2019), "Bayesian Updating Rules and AGM Belief Revision." *Journal of Economic Theory*, 179, 455–475.

Brainard, Jeffrey and Jia You (2018), "What a Massive Database of Retracted Papers Reveals about Science Publishing's 'Death Penality'." *Science*.

Funk, Cary and Brian Kennedy (2016), "The New Food Fights: U.S. Public Divides Over Food Science." *Pew Research*.

Galperti, Simone (2019), "Persuasion: The Art of Changing Worldviews." *American Economic Review*, 109, 996–1031.

Gilboa, Itzhak, Andrew Postlewaite, Larry Samuelson, and David Schmeidler (2019), "What are Axiomatizations Good For?" *Theory and Decision*, 86, 339–359.

Gilboa, Itzhak, Andrew Postlewaite, and David Schmeidler (2009), "Is it Always Rational to Satisfy Savage's Axioms?" *Economics and Philosophy*, 25, 285–296.

Harris, Paul and Melissa Koenig (2006), "Trust in Testimony: How Children Learn about Science and Religion." *Child Development*, 77, 505–524.

Kahan, Dan (2013), "Ideology, Motivated Reasoning, and Cognitive Reflection." *Judgment and Decision Making*, 8, 407–424.

Karni, Edi and Marie-Louise Viero (2013), ""Reverse Bayesianism:" A Choice-Based Theory of Growing Awareness." *American Economic Review*, 103, 2790–2810.

Lipman, Barton (1999), "Decision Theory Without Logical Omniscience: Toward an Axiomatic Framework for Bounded Rationality." *Review of Economic Studies*, 66, 339–361.

Mettler, Katie (2017), "How a Lab Chemist Went from 'Superwoman' to Disgraced Saboteur of more than 20,000 Drug Cases." *Washington Post*.

Nathanson, Melvyn (2008), "Desperately Seeking Mathematical Truth." *Notices American Mathematical Society*, 55, 773.

Nickerson, Raymond (1998), "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises." *Review of General Psychology*, 2, 175–220.

Ortoleva, Pietro (2012), "Modeling the Change of Paradigm: Non-Bayesian Reaction to Unexpected News." *American Economic Review*, 102, 2410–2436.

Popper, Karl (1963), *Conjectures and Refutations*. Routledge.

Rabin, Matthew and Joel Schrag (1999), "First Impressions Matter: A Model of Confirmatory Bias." *Quarterly Journal of Economics*, 114, 37–82.

Radner, Roy (2002), "Bayesian Analysis and Model Revision for k'th Order Markov Chains with Unknown k." Working Paper.

Rubinstein, Ariel and Yuval Salant (2006), "A Model of Choice from Lists." *Theoretical Economics*, 1, 3–17.

Sadler, Evan (2020), "Influence Campaigns." Working Paper.

Savage, Leonard (1954), *The Foundations of Statistics*, second edition. Dover Publications Inc. 2nd Ed. published 1972.

Stoye, J. (2011), "Statistical Decisions under Ambiguity." *Theory and Decision*, 70, 129–148.

Tappin, Ben, Leslie van der Leer, and Ryan McKay (2017), "The Heart Trumps the Head: Desirability Bias in Political Belief Revision." *Journal of Experimental Psychology: General*, 146, 1143–1149.

Wilson, Andrea (2014), "Bounded Memory and Biases in Information Processing." *Econometrica*, 82, 2257–2294.