Learning from a Black Box

Shaowei Ke* Brian Wu† Chen Zhao‡§

July 2021

Please click here for the newest version.

Abstract

We study a decision maker's learning behavior when she receives recommendations from a black box, i.e., the decision maker does not understand how the recommendations are generated. We introduce four reasonable axioms and show that they cannot be satisfied simultaneously. We analyze various relaxations of the axioms. In one relaxation, we introduce and characterize an updating rule, the contraction rule, which has two parameters that map each recommendation to a recommended belief and the trustworthiness of the recommendation, respectively. The decision maker's posterior is formed by mixing her prior with the recommended belief according to the trustworthiness measure.

*Department of Economics, University of Michigan. Email: shaoweik@umich.edu.

†Strategy Department, Stephen M. Ross School of Business, University of Michigan. Email: wux@umich.edu.

‡Faculty of Business and Economics, University of Hong Kong. Email: czhao@hku.hk.

§We are grateful to Tilman Börgers, Jon Eguia, Heng Liu, David Miller, Harry Pei, Zhaoran Wang, and seminar participants at UMich and Toronto for helpful discussions.

1 Introduction

It is becoming increasingly common that people make decisions with the help of complex machine learning algorithms. For example, we often make decisions based on recommendations from online marketplaces and online content platforms. After Deepmind's AlphaGo became the first computer program to defeat professional Go players, many Go players started to take recommendations from machine learning computer programs such as KataGo and Leela when playing or practicing.

Such recommendations are often generated based on datasets with hundreds of billions of variables and algorithms with hundreds of millions of parameters, and it is nearly impossible for people to understand how the recommendations are generated. Indeed, not even the programmers of the algorithms themselves understand what the algorithm has learned from the data, what the internal logic of the mapping from the input to the output is, and how exactly the algorithm manages to make good recommendations. Even if a programmer claims to understand how the algorithm works, it is unlikely that she can explain the rationale or theory behind the recommendations to the decision makers who will make choices based on those recommendations.[1]

For these reasons, a complex machine learning algorithm is often called a *black box* (BB). In this paper, we will use this terminology more broadly. If the decision maker does not understand how a recommender generates its recommendations, we will call such a recommender a BB. For example, if a decision maker receives recommendations from an expert, but the decision maker does not understand (probabilistically) how the expert comes up with the recommendations, we will call the expert a BB.

We are interested in studying a decision maker's learning behavior when she receives recommendations from a BB whose recommendations are typically, though not always, quite accurate and whose ability to make recommendations does not change in a predictable way

[1] A growing literature studies how to make complex machine learning algorithms more interpretable and explainable. See Guidotti, Monreale, Ruggieri, Turini, Giannotti, and Pedreschi (2018) for a recent survey.

over time. For example, such a BB may be a complex machine learning computer program whose training dataset is fixed but sufficiently large to generate good (but imperfect) recommendations.

Specifically, in each period, the decision maker faces a set of actions to choose from, and the BB recommends one. The decision maker has a prior belief over states of the world, but she does not understand how the BB's recommendations are generated. That is, she does not know the conditional distribution of the recommendation given each state to perform Bayesian updating.

If not Bayesian updating, how does the decision maker process the BB's recommendation? Recall that the BB's recommendation is often correct. When it is correct, it means that the recommended action is the best for the decision maker. Therefore, from the decision maker's point of view, the BB essentially recommends the *set of beliefs* over states of the world under which the recommended action is optimal. The primitive of our theory, the decision maker's updating rule, takes the decision maker's prior belief and a set of recommended beliefs and maps them to a posterior belief.

We introduce four axioms imposed on the updating rule. Below, we describe them in a special case in which there are only two states, the high state and the low state. The first axiom is monotonicity. It says that if a prior p puts more weight on the high state than another prior q does, p's posterior should still put more weight on the high state than q's posterior upon receiving the same recommendation. The second axiom is partial obedience. It is based on the assumption that the BB is often accurate. It states that for any recommendation, there is always some prior belief such that the decision maker will follow the BB's recommendation, even though her prior does not agree with the recommendation. The third axiom, sensitivity to repetition, stems from the assumption that the BB's recommendation is imperfect. It requires that if the decision maker's prior is inconsistent with the BB's recommendation, she should be convinced gradually as the BB repeats this recommendation. The idea of the last axiom, regularity, is taken from Bayes' rule. Roughly speaking, it assumes

that the decision maker complies with the black box in the long run upon receiving repeated recommendations.

Our first main result shows that there does not exist any updating rule that satisfies the four axioms simultaneously; every updating rule must violate at least one of the four axioms. Therefore, our approach follows the spirit of the classic impossibility theorem by Arrow (1951). We introduce a set of reasonable axioms to characterize how the decision maker incorporates the new information BB provides to update her belief. Each individual axiom is plausible, reflecting the strengths of the BB in often offering good, albeit imperfect, recommendations. Collectively, however, these axioms generate a negative result, meaning that the decision maker is bound to face a trade-off between some desirable properties when she does not understand how the recommendations are generated. This points to an challenging question–how can a decision maker utilizes the often-correct recommendations that are developed from enormous data and advanced algorithms, yet still eventually reach the truth in an internally coherent way (see a related discussion in Kahneman, Sibony, and Sunstein (2021))?

Similar to how the rich literature expands Arrow's (1951) pioneering theorem, our model moves beyond identifying the negative result, as our ultimate objective is to understand how to mitigate the trade-off caused by using the BB. We investigate whether we should modify the axioms so that we can find avenues to mitigate the trade-off—that is, the decision makers can both leverage BB's strengths while avoiding its limitations, as much as possible.

Therefore, we analyze several relaxations of the axioms. We first focus on regularity. Regularity requires that if the BB recommends several sets of beliefs $I_1, I_2, \ldots, I_n$ repeatedly in some alternating way, the decision maker's asympototic belief should be sufficiently close to the intersection of $I_1, I_2, \ldots, I_n$. We weaken it by requiring that her asymptotic belief be sufficiently close to one recommended set of beliefs I when the BB recommends I and only I repeatedly. We show that this weakening, together with sensitivity of repetition and an appropriate version of monotonicity, characterizes an updating rule called the contraction

rule.

The contraction rule has two parameters. One is a function that maps each recommendation to one recommended belief. The other is a function that maps each recommendation to a measure of how much the decision maker trusts the recommendation. The decision maker's posterior is given by mixing her prior with the recommended belief, weighted by the measure of trustworthiness.

Finally, we relax some assumptions of our theory to incorporate the idea of bounded memory. The fact that the updating rule does not directly depend on past recommendations is not an issue in the Bayesian benchmark. The joint prior in the Bayesian benchmark is a sufficient statistic of past information. However, this assumption may be restrictive in our learning problem. Therefore, we assume that the updating rule may depend on past recommendations. This leads to natural weakening of our axioms except for regularity. We find that in contrast to the case in which we relax regularity and characterize the contraction rule, weakening the other axioms by introducing some history dependence does not help us bypass the negative results.

1.1 Related Literature

It is well recognized in the machine learning literature that relying on BBs to make decisions may cause biases (Pedreshi, Ruggieri, and Turini (2008); Barocas and Selbst (2016)); legal liability issues (Kingston (2016); Bathaee (2018)); and severe consequences (Wexler (2017); Nunes, Reimer, and Coughlin (2018)). To open the BB, the literature follows two directions: (i) ex ante designing interpretable models to make predictions (see, for example, Doshi-Velez and Kim (2017) and, in an economic context, Ke, Zhao, Wang, and Hsieh (2021)); (ii) ex post seeking to explain the predictions made by BBs (see, for example, Ribeiro, Singh, and Guestrin (2016) and Guidotti et al. (2018)). In this paper, we keep the BB closed but investigate how a decision maker incorporates its recommendations into her beliefs. Our results add to the literature by highlighting the general difficulty of learning from BBs:

The decision maker faces a trade-off between some desirable properties since she does not understand how the recommendations are generated.

In our setup, new information comes in as a set of beliefs that are consistent with the BB's recommendation. Thus, our updating rule takes the decision maker's prior and a subset of probability measures as input and returns a posterior. In the decision theory literature, Zhao (2021) and Dominiak, Kovach, and Tserenjigmid (2021) consider similar primitives and propose updating rules that select, from the given set of distributions, the posterior belief closest to her prior according to some subjective divergence measure. Both papers interpret the subset of probability measures as a constraint with which the decision maker's posterior must be consistent. In information theory, there is also a literature that considers belief updating when new information imposes constraints on the probability distribution; see, for example, Shore and Johnson (1980); Skilling (1988); and Caticha (2004). In contrast to all the papers above, we allow the decision maker to not fully trust the BB, and thus her posterior may be outside the set of distributions consistent with its recommendation.

Chambers and Hayashi (2010) and Damiano (2006) model how a decision maker selects her subjective belief from a set of objectively possible probability measures. In these studies, the decision maker does not have a prior to begin with, and thus the selection rule only depends on the probability-possibility set. Ahn (2008) and Gajdos, Hayashi, Tallon, and Vergnaud (2008) take probability-possiblity sets as the primitive to generate ambiguity.

There is a much larger literature on non-Bayesian updating with standard information, i.e., the occurrence of an event. For behavioral models, see, for example, Rabin and Schrag (1999); Rabin (2002); Mullainathan, Schwartzstein, and Shleifer (2008); and Gennaioli and Shleifer (2010). For decision-theoretic models, see, for example, Epstein (2006); Ortoleva (2012); Zhao (2020); and Kovach (2021). Of these studies, Epstein (2006) and Kovach (2021) characterize the prior-biased updating rule: The decision maker's behavioral posterior is a convex combination of her prior and the Bayesian posterior. Our contraction rule also features a convex combination between the decision maker's prior and the recommended

belief, but is defined on a completely different primitive.

2 The Binary Case

There is a binary state $\theta \in \Theta = \{0, 1\}$, and the set of all possible actions is $\mathcal{A}$. In each period (of finitely or infinitely many periods), the decision maker faces a nonempty finite set of actions, which is a subset of $\mathcal{A}$ denoted by A, B, C. Facing a finite set of actions A, the decision maker needs to choose an action $a \in A$. The decision maker never observes θ, but she has a subjective probabilistic assessment of $\theta = 1$, denoted by p, q, r. If her belief about $\theta = 1$ is $p \in \Delta(\Theta) = [0, 1]$, her expected utility of action a is $U(a, p) = pU(a, 1) + (1 - p)U(a, 0)$.

The decision maker receives a recommendation in each period, but the recommendation does not come from a Bayesian expert. If it comes from a Bayesian expert, the decision maker will learn the true state in the long run under standard assumptions with no difficulty.[2] In our model, first, the decision maker does not know how the recommendation is generated. In other words, the decision maker does not know the conditional distribution of the recommendation given each state, which means that she does not have a joint prior needed to perform Bayesian updating after receiving the recommendation. Second, the decision maker understands that the recommendation is quite accurate, although not always correct, and the quality of the recommendation is not changing over time. These assumptions will become important as we interpret the axioms to be imposed on the decision maker's learning behavior. We call such a recommender a BB.

The BB may, for instance, be a complex machine learning algorithm. It is often the case that no one, including the designer of the algorithm, understands how the input is translated into the recommendation exactly. In the meantime, the decision maker understands that the algorithm often, though not always, makes correct recommendations given enough data. Note that we focus on the case in which the quality of the recommendation is constant over

[2]For example, the Bayesian expert has the same objective function as the decision maker, and reveals a signal that correlates with the true state in each period, and the decision maker knows the joint distribution of the state and the signals.

time. In this example, this means that we consider the situation in which the algorithm's training data remain unchanged over time.

Then how does the decision maker interpret a BB's recommendation? Recall that the BB often makes correct recommendations. When a recommendation $a \in A$ is correct, it means that the decision maker's utility of a is higher than any $b \in A$, or equivalently, the probability that $\theta = 1$ should be in the set $I(a, A) = \{p \in [0,1] | U(a,p) \geq U(b,p) \text{ for any } b \in A\}$. Note that this does not suggest that the BB consciously knows what the decision maker's utility function is. It is merely the consequence of assuming that the BB's recommendation is correct.

Because U is an expected utility function, it must be the case that $I(a, A)$ is a closed interval in $[0,1]$ for any $a \in A \in \mathcal{A}$. Let $\mathcal{I} = \{[\alpha, \beta] \subsetneq [0,1] | \alpha < \beta\}$ be the set of all nontrivial closed intervals that are not $[0,1]$. For simplicity, we exclude situations in which the recommendation is uninformative or decisive. Let $a(p, A)$ denote the set of optimal actions in A if the decision maker believes that $\theta = 1$ with probability p. In other words, $a(p, A) = \{a \in A | U(a,p) \geq U(b,p) \text{ for any } b \in A\}$.

The Updating Rule. In each period, given her current belief about $\theta = 1$ and what the BB suggests from her point of view ($I(a, A)$ if the BB recommends a from A), the decision maker forms a new belief about $\theta = 1$. Therefore, the decision maker's *updating rule* is a function $\pi : [0,1] \times \mathcal{I} \to [0,1]$.

This definition allows the decision maker to not fully trust the BB, because $\pi(p, I)$ may not be in I. Second, it assumes that in each period, the decision maker applies new information $I \in \mathcal{I}$ to her current belief using the same π. Last, the decision maker's posterior belief about $\theta = 1$ only depends on the current belief and the current new information, and does not depend on past information. The last assumption will be relaxed in Section 5.

To simplify notation, we write p_I instead of $\pi(p, I)$. Recursively, we define $p_{I_1 I_2 \ldots I_n} = \pi(p_{I_1 I_2 \ldots I_{n-1}}, I_n)$. In other words, $p_{I_1 I_2 \ldots I_n}$ is the decision maker's posterior after learning $I_1, I_2, \ldots, I_n$ sequentially. For any $I \in \mathcal{I}$ and $n \in \mathbb{N}$, let I^n denote a string of n consecutive

I's. Then, for example, $p_{I^2} = p_{II}$, $p_{I^3} = p_{III}$. Furthermore, $(I^m J^n)^k$ denotes a string of k consecutive $I^m J^n$'s. For example, $p_{(I^2J)^2} = p_{IIJIIJ}$.

Richness. We assume that the pair $(U, \mathcal{A})$ is *rich*: For any $\alpha, \beta \in [-1, 1]$, there exists an action $a \in \mathcal{A}$ such that $U(a, 0) = \alpha$ and $U(a, 1) = \beta$.

Richness immediately implies that for any $I \in \mathcal{I}$, there exists some $a \in A \in \mathcal{A}$ such that $I(a, A) = I$. In other words, for any nontrivial closed interval I that is not $[0, 1]$, there exist some set of actions and a recommended action such that the information suggested by the recommendation is exactly I. Henceforth, every element of $\mathcal{I}$ will be called a *recommendation*.

Axioms. We do not make specific functional-form assumptions about the updating rule. Rather, we impose some reasonable axioms on it. The first axiom is monotonicity, which says that fixing any recommendation I, the decision maker's posterior belief about $\theta = 1$ is increasing in her prior belief.

Axiom 1 (Monotonicity)**.** *For any $p, q \in \Delta(\Theta)$ and $I \in \mathcal{I}$, $p \geq q$ implies $p_I \geq q_I$.*

The next axiom relies on the implicit assumption that the BB's recommendation is quite accurate.

Axiom 2 (Partial Obedience)**.** *For any $I \in \mathcal{I}$, there exists $p \notin I$ such that $p_I \in I$.*

This axiom says that at least under some prior belief, the decision maker will simply follow the BB's recommendation. We can equivalently describe this axiom in terms of recommended actions: For any $a \in A \in \mathcal{A}$, there exists some $p \in [0, 1]$ such that $a \notin a(p, A)$ but $a \in a(p_{I(a,A)}, A)$. In other words, there are always some cases in which the decision maker's optimal action before receiving the BB's recommendation differs from the BB's recommendation, but after receiving the recommendation the decision maker changes her mind and follows the recommendation.

Note that partial obedience allows the decision maker to always be indifferent between the recommended action and some other actions. In other words, the decision maker does not have to believe that the recommended action is strictly the best.

Our next axiom is based on the following observation. Recall that the BB often makes accurate recommendations but not always. Therefore, one would expect that if the BB makes the same recommendation repeatedly, the decision maker's posterior would respond to the repetition. For example, suppose that the decision maker's initial belief about $\theta = 1$ is .1, but the BB suggests $[.5, 1]$. The first time she sees this recommendation, the decision maker may move her belief toward the interval to some extent—but if she receives this recommendation repeatedly, she may eventually move her belief to somewhere close to .75.[3] In other words, seeing the recommendation once and seeing it many times should lead to different posteriors, especially when the decision maker's initial belief is inconsistent with the BB's recommendation.

Axiom 3 (Sensitivity to Repetition)**.** *For any $p \in \Delta(\Theta)$ and $I \in \mathcal{I}$, if $p \notin I$ then there exist $m, n \geq 1$ such that $m \neq n$ and $p_{I^m} \neq p_{I^n}$.*

To state our last axiom, we first define a special sequence of recommendations.

Definition 1. *A sequence of recommendations $\{I_n\}_{n=1}^{\infty}$ is reinforcing if there exists $N \geq 1$ and $I, J \in \mathcal{I}$ with $I \cap J \neq \emptyset$ such that $I_{2kN+i} = I$ and $I_{(2k+1)N+i} = J$ for any $k \geq 0$ and $1 \leq i \leq N$.*

Consider an example in which the decision maker faces the set of actions $A = \{a_1, a_2, a_3\}$ in all odd periods and $B = \{b_1, b_2\}$ in all even periods. Suppose that the BB always recommends a_2 from A and b_1 from B, and $I(a_2, A) = [.8, 1]$ and $I(b_1, B) = [.7, .9]$. This is a reinforcing sequence of recommendations. One way to interpret it is that perhaps what the BB really wants to recommend is the intersection of those intervals, $[.8, .9]$, but the sets of actions do not allow the BB to do that. Therefore, it seems reasonable to require that in

[3] By contrast, if the BB is always correct, the decision maker's belief may move to .75 the first time she receives the recommendation $[.5, 1]$ and stay there, as the BB recommends $[.5, 1]$ repeatedly.

this case, the decision maker's long-run belief will be arbitrarily close to $[.8, .9]$. The axiom below captures this idea.

Axiom 4 (Regularity)**.** *For any $p \in \Delta(\Theta)$, if $\{I_n\}_{n=1}^{\infty}$ is reinforcing then each accumulation point of $\{p_{I_1 I_2 \ldots I_n}\}_{n=1}^{\infty}$ is in $\bigcap_{n=1}^{\infty} I_n$.*

Note that in each period, the decision maker's posterior only depends on her prior and the current recommendation. It may appear that the decision maker should not be able understand how to combine all the recommendations that she has received. This is not true, because the decision maker's prior could represent her summary of the past recommendations if the updating rule satisfies certain invertibility conditions, which may allow her to combine all recommendations indirectly.

Before introducing our first main result, let us consider several natural updating rules and discuss which axioms they satisfy. Let $d(\cdot, \cdot)$ be the Euclidean metric.

1. Consider the updating rule

$$\pi(p, I) = \operatorname*{argmin}_{q \in I} d(p, q).$$

   This updating rule is often used in misspecified learning models. Under this updating rule, the decision maker fully trusts the BB, because whatever her prior is, her posterior is always consistent with the BB's recommendation. This updating rule satisfies all the axioms except sensitivity to repetition.

2. Consider the updating rule

$$\pi(p, I) = \varepsilon p + (1 - \varepsilon) \operatorname*{argmin}_{q \in I} d(p, q)$$

   with $\varepsilon \in (0, 1)$. Recall that sensitivity to repetition captures the idea that the BB is not always correct. Therefore, a natural attempt to fix the previous updating rule so

that all the axioms may hold is to allow the decision maker to not fully trust the BB. The parameter ε is introduced to capture this. However, this updating rule satisfies all the axioms except partial obedience. The violation of partial obedience is descriptively unrealistic, because it implies that there can be some recommendation such that no matter how close the decision maker's prior is to the recommendation, she never takes the recommended action.

3. Consider the updating rule

$$\pi(p, [\alpha, \beta]) = \varepsilon p + (1 - \varepsilon)\frac{\alpha + \beta}{2}$$

with $\varepsilon \in (0, 1)$. One issue with the previous updating rule is that given a recommendation I, conditional on its being correct, the decision maker is only willing to move her posterior to the closest boundary of I. Therefore, the fact that the decision maker does not believe that the BB is always correct brings her posterior out of I. This updating rule fixes this issue, because conditional on the recommendation I being correct, the decision maker moves her posterior to the center of I. However, it can be verified that this updating rule satisfies all the axioms except regularity.

4. One may wonder if there is any updating rule that satisfies all the axioms except monotonicity. Here is one such example:

$$\pi(p, I) = \varepsilon p + (1 - \varepsilon) \operatorname*{argmin}_{q \in I} d(p, q)$$

with $\varepsilon \in (-1, 0)$. This is an updating rule in which the decision maker seems to overly trust the BB, and it satisfies all the axioms except monotonicity.

From these examples, one can see that it seems difficult to find an updating rule that satisfies all the axioms simultaneously. Our first main result below shows that this is indeed true in general.

Theorem 1. *There does not exist any updating rule that satisfies monotonicity, partial obedience, sensitivity to repetition, and regularity.*

The theorem shows that learning from a BB must violate at least one of the axioms introduced previously. In fact, as will be shown in Proposition 2 in Section 4, such violation does not depend on the decision maker's initial prior.

To illustrate the proof idea, suppose the decision maker has prior p with $p < \alpha < 1$. Let $I = [\alpha, 1]$ and $J = [0, \alpha]$. The goal is to construct a belief q such that starting from q, if the decision maker learns a reinforcing sequence of I, J's, her belief does not converge to α. First, the interaction of partial obedience, regularity, and monotonicity ensures that after learning I a number of times, the decision maker's posterior must go strictly beyond α. This is because regularity requires that the decision maker's belief be higher than the "obedient" prior (i.e., the prior described in partial obedience) if she learns a long enough sequence of I's. Therefore, by monotonicity, her belief has to jump into I if she learns I one more time. Then sensitivity to repetition implies that the decision maker's belief cannot stay at α forever, and thus must eventually jump to some level $r > \alpha$. Let q be the belief right before the jump from $[0, \alpha]$ into $(\alpha, 1]$. The second step is to let the decision maker learn I and J alternately starting from belief q. Suppose that after one iteration of I, J, the decision maker's belief, q_{IJ}, is weakly higher than q. Then, by monotonicity, if she learns one more I, her belief will be weakly higher than r. In fact, monotonicity can be applied inductively to show that every time the decision maker learns I, her belief will be weakly higher than r. Now suppose that after one iteration of I, J, the decision maker's belief, q_{IJ}, is strictly lower than q. Similarly, by applying monotonicity inductively, one can show that every time the decision maker learns J, her belief will be weakly lower than q_{IJ}. Hence, in both cases, the decision maker's belief cannot converge to α, which establishes the impossibility result.

3 The General Case

The impossibility result in the binary case can be generalized. Let the state space Θ be a compact metric space, and Σ be the Borel σ-algebra defined on Θ. Let $\Delta(\Theta)$ be the set of all Borel probability measures defined on Θ. We endow $\Delta(\Theta)$ with the topology of weak convergence. Again, let p, q, r be generic elements of $\Delta(\Theta)$.

For any $a \in \mathcal{A}$, the decision maker's expected utility of a is $U(a,p) = \int_\Theta u(a,\theta)p(d\theta)$. To ensure that the expected utility function is well defined, we assume that $u(a,\cdot)$ is continuous and bounded for each $a \in \mathcal{A}$.

Define $I(a,A)$ in the same way for any nonempty finite $A \subseteq \mathcal{A}$ and $a \in A$. It is clear that $I(a,A) \subseteq \Delta(\Theta)$ is defined by $|A|-1$ linear inequalities: Each inequality is given by $U(a,p) \geq U(b,p)$ for some $b \in A$. Thus, $I(a,A)$ must be convex.

We call a subset $H \subseteq \Delta(\Theta)$ a *probabilistic half-space* if there exists a continuous and bounded function $\tilde{u}$ such that

$$H = \left\{ p \in \Delta(\Theta) \middle| \int_\Theta \tilde{u} dp \geq 0 \right\}.$$

We call a subset $I \subseteq \Delta(\Theta)$ a *probabilistic polytope* if it is the intersection of a finite set of probabilistic half-spaces. We say that a probabilistic polytope I is *nontrivial* if $\text{int}(I)$ is neither $\emptyset$ nor $\Delta(\Theta)$. Let $\mathcal{I}$ denote the set of all nontrivial probabilistic polytopes. In other words, we assume that each recommendation provided by BB is neither uninformative nor good enough to point to a zero-measure set of beliefs.

This setup has several interesting special cases.

1. The state space $\Theta = \{1, 2, \ldots, n\}$ is finite. In this case, the decision maker's belief is one point in the simplex of $\mathbb{R}^n$ and a recommendation $I(a,A)$ is a convex polytope, with each point in it denoting one belief in the simplex of $\mathbb{R}^n$ the decision maker may take into account when updating her belief.

2. The state space is $\Theta = \Delta(\{0,1\})$. In this case, the decision maker holds a second-order belief over two actual states, 0 and 1. In other words, her belief is a distribution over all possible probabilities that the state is 1. The decision maker understands that her prior second-order belief may be incorrect. When she updates, she takes into account other second-order beliefs that are suggested by the BB's recommendation.

Remark 1. *In the second special case, note that if the BB's recommendation is something the decision maker understands probabilistically, as in some Bayesian benchmark, the BB's recommendation should be an event rather than a set of beliefs. This is the key ingredient of our setup that makes the recommender a BB.*

Richness. We assume that the pair $(U, \mathcal{A})$ is *rich*: For any continuous and bounded function $f : \Theta \to [-1, 1]$, there exists an action $a \in \mathcal{A}$ such that $u(a, \cdot) = f$.

As before, this richness assumption implies that for any probabilistic polytope $I \in \mathcal{I}$, there exist some set of actions and a recommended action such that the information suggested by the recommendation is exactly I. Again, every element of $\mathcal{I}$ will be called a recommendation henceforth.

Lemma 1. *If $(U, \mathcal{A})$ is rich, then for any $I \in \mathcal{I}$ there exists some $a \in A \subseteq \mathcal{A}$ such that $I(a, A) = I$.*

The Updating Rule. In each period, given her current belief $p \in \Delta(\Theta)$ and what the BB suggests from her point of view ($I(a, A)$ if the BB recommends a from A), the decision maker forms a new belief in $\Delta(\Theta)$. Therefore, the decision maker's *updating rule* is a function $\pi : \Delta(\Theta) \times \mathcal{I} \to \Delta(\Theta)$.

While partial obedience and regularity in the binary case continue to work in the general case, monotonicity and sensitivity to repetition require more work. In the binary case, the space of probability measures is $[0, 1]$. It is linearly ordered so that we may state inequalities such as $p \geq q$. The space of probability measures in the general case is not necessarily linearly ordered. Therefore, monotonicity cannot be applied directly here.

In addition, sensitivity to repetition becomes too weak to be desirable in the general case. To see this, suppose the decision maker initially believes that action b is strictly better than action a. Then she receives the recommendation $I(a, \{a, b\})$ repeatedly. Note that in the general case, there are many distinct beliefs that make a and b equally good. Therefore, sensitivity to repetition can be satisfied in an uninteresting way: After receiving the recommendations, the decision maker always believes that a and b are equally good, but she keeps changing her posteriors from one belief that implies that a is indifferent to b to another such belief.

To adapt monotonicity and sensitivity to repetition to the general case, we first introduce a preorder (reflexive and transitive binary relation) on $\Delta(\Theta)$.

Definition 2. *For any actions $a, b \in \mathcal{A}$, we say that $p \in \Delta(\Theta)$ is less confident than $q \in \Delta(\Theta)$ about a being better than b if $\alpha p + (1-\alpha)r \in I(a, \{a, b\})$ implies $\alpha q + (1-\alpha)r \in I(a, \{a, b\})$ for any $\alpha \in [0, 1]$ and $r \in \Delta(\Theta)$. We denote this by $p \sqsubset_b^a q$.*

The following proposition shows that if $I(a, \{a, b\})$ is nontrivial, then $p \sqsubset_b^a q$ has a simple cardinal representation.

Proposition 1. *For any $p, q \in \Delta(\Theta)$ and $a, b \in \mathcal{A}$ with $I(a, \{a, b\}) \in \mathcal{I}$,*

$$p \sqsubset_b^a q \;\Leftrightarrow\; U(a, p) - U(b, p) \leq U(a, q) - U(b, q).$$

Note that $U(\cdot, p)$ and $U(\cdot, q)$ share the same Bernoulli index; they only differ by the belief. Below are two notions of monotonicity that extend the one in the binary case to the general case.

Axiom 5 (Weak Monotonicity)**.** *For any $I \in \mathcal{I}$, $p \sqsubset_b^a q$ and $I \subseteq I(a, \{a, b\})$ implies $p_I \sqsubset_b^a q_I$.*

Axiom 6 (Strong Monotonicity)**.** *For any $I \in \mathcal{I}$, $p \sqsubset_b^a q$ implies $p_I \sqsubset_b^a q_I$.*

The idea behind these notions of monotonicity is simple. If a prior is more confident than another prior about action a being better than action b, upon receiving the same

recommendation, the former prior should still be more confident than the latter prior about action a being better than action b. Strong monotonicity applies to any recommendation, while weak monotonicity only applies to recommendations that imply a is better than b. We only need the weak version for the impossibility result below. Strong monotonicity will become useful when we analyze how to bypass the impossibility result.

Similarly, we extend the notion of sensitivity to repetition to the general case as follows.

Axiom 7 (Sensitivity to Repetition*)**.** *For any $p \in \Delta(\Theta)$ and $I \in \mathcal{I}$, if $p \notin I(a, \{a, b\}) \supseteq I$ then there exist $m, n \geq 1$ such that $m \neq n$ and $p_{I^m} \not\succeq_b^a p_{I^n}$.*

Thus, if the decision maker initially believes that action b is strictly better than action a, then upon repeatedly receiving recommendation I, which implies that a is better than b, her confidence about a being better than b does not always stay the same.

Theorem 2. *There does not exist any updating rule that satisfies weak monotonicity, partial obedience, sensitivity to repetition*, and regularity.*

On the one hand, this negative result illustrates the tension between several desirable properties of learning from a BB. On the other hand, one may wonder whether the assumptions can be relaxed in a reasonable way to bypass this negative result. Below, we discuss several such attempts.

4 The Contraction Rule

In this section, we explore how weakening regularity could help us avoid the negative results. First, note that the main idea behind regularity comes from Bayes' rule. If a Bayesian decision maker knows that several events have happened, the support of her belief should be equal to the intersection of those events.[4] When the decision maker is learning from a BB, however, such a normatively appealing axiom could be too demanding. Nonetheless,

[4]Regularity per se is not satisfied by Bayes' rule, because our setup is different from a standard Bayesian learning setup. See Remark 1.

the following weakening of regularity does seem like a minimum requirement we want the updating rule to satisfy.

Axiom 8 (Weak Regularity)**.** *For any $p \in \Delta(\Theta)$ and $I \in \mathcal{I}$, each accumulation point of $\{p_{I^n}\}_{n=1}^{\infty}$ is in I.*

To illustrate what is lost from weakening regularity, consider the following type of recommendation sequence.

Definition 3. *A sequence of recommendations $\{I_n\}_{n=1}^{\infty}$ reveals the equivalence of $a, b \in \mathcal{A}$ if it is a reinforcing sequence consisting of $I(a, \{a, b\})$ and $I(b, \{a, b\})$.*

The following proposition shows that with weak regularity, it is generally hard to learn the equivalence between actions a and b even if BB recommends each of them alternately from the binary menu $\{a, b\}$. In addition, the proposition implies that the negative result in Theorem 2 holds regardless of the decision maker's initial prior.

Proposition 2. *Given any updating rule that satisfies weak monotonicity, partial obedience, sensitivity to repetition*, and weak regularity, for any $p \in \Delta(\Theta)$ and any $a, b \in \mathcal{A}$ such that $I(a, \{a, b\}), I(b, \{a, b\}) \in \mathcal{I}$, there exists $\{I_n\}_{n=1}^{\infty}$ that reveals the equivalence of a, b but $U(a, p_{I_1 I_2 \ldots I_n}) - U(b, p_{I_1 I_2 \ldots I_n})$ does not converge to 0.*

Next, we define a generalization of the updating rule introduced in Example 3 in Section 2. An updating rule $\pi : \Delta(\Theta) \times \mathcal{I} \to \Delta(\Theta)$ is a *contraction rule* if there exists a mapping $\rho : \mathcal{I} \to \Delta(\Theta)$ such that $\rho(I) \subseteq I$ for all $I \in \mathcal{I}$, and a functional $\varepsilon : \mathcal{I} \to (0, 1)$, such that

$$\pi(p, I) = \varepsilon(I) \cdot p + (1 - \varepsilon(I)) \cdot \rho(I)$$

for any $p \in \Delta(\Theta)$ and $I \in \mathcal{I}$.

A contraction rule says that from the decision maker's point of view, each recommendation I can be reduced to one belief in I, $\rho(I)$, and a measure of how much the decision maker

trusts this recommendation, $\varepsilon(I)$. The decision maker's posterior is formed by mixing her prior with $\rho(I)$ with probability $\varepsilon(I)$.

The next result shows that weak regularity helps us avoid the impossibility results and obtain the contraction rule as a representation of the updating rule.

Theorem 3. *Suppose $|\Theta| \geq 3$. An updating rule satisfies strong monotonicity, sensitivity to repetition*, and weak regularity if and only if it is a contraction rule. Furthermore, a contraction rule (ρ, ε) satisfies partial obedience if $\rho(I) \in \operatorname{int}(I)$ for all $I \in \mathcal{I}$.*

The key step in the proof is to exploit strong monotonicity to show that the difference between the posteriors, as a finite signed measure, has to be a scalar multiple of the difference between the priors. In other words, in the vector space of finite signed measures (denoted as Δ^*), for any two priors p and q, the line connecting them must be parallel to the line connecting their posteriors p_I and q_I. Thus, the role of strong monotonicity in our theory is analogous to that of the independence axiom in expected utility theory, which only exerts its full power when there are at least three states. Furthermore, the parallel property implies that the updating rule is weakly continuous. This, together with the compactness of Θ, the properties of Δ^*, and the other axioms, allows us to apply the Schauder–Tychonoff fixed point theorem and show that the updating rule has a fixed point in I (denoted as r). Thus, for any p, the line connecting p and r must be parallel to the line connecting p_I and r. Then sensitivity to repetition* ensures that the updating rule is indeed a contraction rule.

5 Bounded Memory

In our setup, the decision maker's posterior only depends on the current prior and recommendation but not on past recommendations. To put this differently, we assume that the decision maker treats her prior as a sufficient statistic of what she has learned in the past. In Bayesian updating, the decision maker's joint prior indeed summarizes all the information needed to interpret the new signal, including what she learned from the past. However,

since our decision maker does not have a joint prior to perform Bayesian updating, one may suspect that allowing her posterior to depend on past recommendations could be useful.

To investigate this issue, we assume that the decision maker can remember the last $M \geq 1$ recommendations—that is, the decision maker has a bounded memory. We do not consider the case in which $M = +\infty$ because it is not realistic and the updating rule that can accommodate memories with arbitrary length is so flexible that some of our previous results would make little sense in that setup.[5]

The result in this section can be extended to the general case with the techniques in Section 3, but for ease of presentation, we present it in the binary case. Given her current belief and the M most recent recommendations, the decision maker forms a new belief. The set of all possible recommendations is $\mathcal{I} = \{[\alpha, \beta] \subsetneq [0,1] | 0 \leq \alpha < \beta \leq 1]\} \cup \{\emptyset\}$, with $\emptyset$ (no recommendation) used only when the decision maker has received fewer than M recommendations. Thus, her information set before updating is an element of

$$\mathbb{I} = \{(I_1, I_2, \ldots, I_M) \in \mathcal{I}^M | I_M \neq \emptyset \text{ and if } I_N = \emptyset \text{ then } I_n = \emptyset \text{ for all } n \leq N\}.$$

Then the decision maker's *updating rule* is a function $\pi : [0,1] \times \mathbb{I} \to [0,1]$.

Let $p_{\langle I_1 \ldots I_{M-1} \rangle I_M} := \pi(p, I_1, I_2, \ldots, I_M)$ with p being the current belief and I_M being the current recommendation. Let $p_{\langle I_1 I_2 \ldots I_{M-1} \rangle I_M I_{M+1} \ldots I_N}$ denote the decision maker's posterior if she has belief p and updates based on $(I_1, \ldots, I_M)$, $(I_2, \ldots, I_{M+1})$, $\ldots$, $(I_{N-M+1}, \ldots, I_N)$, sequentially.

Now we turn to the axioms. Note that monotonicity, in its original form, seems too strong in the bounded memory setting. Consider two situations. In the first, the decision maker's current belief is .5, which is the consequence of learning from the recommendation $[0, .5]$ for several times. In the second, the decision maker's current belief is also .5, but this is the consequence of learning from the recommendation $[.5, 1]$ for several times. Of course, the

[5]For example, Proposition 2 may fail trivially. This is because when the updating rule can depend on any history of recommendations, if we fix an initial prior and a sequence of recommendations, most of our axioms, such as weak monotonicity, impose few restrictions on the sequence of posteriors.

decision maker's initial priors before receiving these recommendations are different in the two situations. Now, suppose the decision maker receives $[.5, 1]$. It seems natural that the decision maker's posteriors should be different in these two situations. However, monotonicity is then violated, because in the last period of both situations, the decision maker's beliefs and the recommendations are identical. The direct translation of monotonicity into the bounded memory setting implies that the decision maker's posteriors should also be identical in both situations, but our discussion appears to suggest otherwise.

Similar examples can be constructed for partial obedience and sensitivity to repetition. Therefore, we extend the main axioms from the previous sections to this new setup with bounded memory by accounting for the possibility that past recommendations may affect current updating.

Axiom 9 (Monotonicity†)**.** *For any $p, q \in [0,1]$ and $(I_1, I_2, \ldots, I_M) \in \mathbb{I}$, $p \geq q$ implies $p_{\langle I_1 I_2 \ldots I_{M-1}\rangle I_M} \geq q_{\langle I_1 I_2 \ldots I_{M-1}\rangle I_M}$.*

Monotonicity† states that given the same history $(I_1, I_2, \ldots, I_{M-1})$, the decision maker's posterior belief after receiving a new recommendation is increasing in her current belief. Under the new monotonicity condition, we cannot conclude that the decision maker's posteriors should be identical in both situations in the previous example, since the history of recommendations are different across the two situations. Therefore, monotonicity† is weaker than in previous sections.

Similarly, partial obedience† and sensitivity† to repetition are weaker than their counterparts in previous sections.

Axiom 10 (Partial Obedience†)**.** *For any $(I_1, I_2, \ldots, I_M) \in \mathbb{I}$, there exists some $p \notin I_M$ such that $p_{\langle I_1 I_2 \ldots I_{M-1}\rangle I_M} \in I_M$.*

Axiom 11 (Sensitivity to Repetition†)**.** *For any $(I_1, I_2, \ldots, I_M) \in \mathbb{I}$ and $p \notin I_M$, there exist $1 \leq m, n \in \mathbb{N}$ such that $m \neq n$ and $p_{\langle I_1 I_2 \ldots I_{M-1}\rangle I_M^m} \neq p_{\langle I_1 I_2 \ldots I_{M-1}\rangle I_M^n}$.*

Since regularity is asymptotic property that is not unaffected by bounded memory, Regularity† is essentially equivalent to the one in the previous sections.

Axiom 12 (Regularity†)**.** *For any $p \in [0,1]$, if $\{I_n\}_{n=1}^{\infty}$ is reinforcing then each accumulation point of $\{p_{\langle I_1 I_2 \ldots I_{M-1}\rangle I_M I_{M+1} \ldots I_n}\}_{n=M}^{\infty}$ is in $\bigcap_{n=1}^{\infty} I_n$.*

The next result shows that even if the decision maker can take past recommendations into consideration, the same negative result applies.

Theorem 4. *There does not exist any updating rule that satisfies monotonicity†, partial obedience†, sensitivity to repetition†, and regularity†.*

6 Conclusion

In this paper, we study a decision maker's learning behavior when she learns from a BB. A BB may be a complicated machine learning algorithm using high-dimensional datasets, or an expert whose process for generating recommendations is not understood by the decision maker. Specifically, the decision maker does not know the conditional distribution of the recommendations given each state, and therefore cannot perform Bayesian updating.

We introduce several reasonable axioms imposed on the decision maker's updating rule, and show that every updating rule must violate at least one of the axioms. Hence, learning from a BB is bound to face a trade-off between the desirable properties described in the axioms.

We examine how we may avoid the negative result. In one of our attempts, we relax an axiom called regularity and find that the weakened regularity, together with some other axioms, lead to an updating rule we call the contraction rule. In the contraction rule, the decision maker reduces each recommendation from the BB to a single recommended belief, and assesses the trustworthiness of each recommendation. When she receives a recommendation, which induces a recommended belief and a measure of trustworthiness, she forms

a posterior by mixing her prior with the recommended belief, weighted by the measure of trustworthiness.

We also use the idea of bounded memory to relax the setup of our theory. That is, we allow the updating rule to depend on past recommendations. Note that Bayes' rule does not rely on past information directly. The joint prior in Bayes' rule is a summary statistic of all past information. Nonetheless, we wonder if past recommendations will play an important role when the decision maker learns from a BB and hence cannot be Bayesian. Allowing the updating rule to depend on past recommendations naturally weakens all the axioms except regularity. In contrast to the case above, in which regularity is relaxed, we find that relaxing the other axioms by allowing them to be more history dependent does not help us bypass the negative result.

References

Ahn, D. S. (2008). Ambiguity without a State Space. *Review of Economic Studies 75*(1), 3–28.

Arrow, K. J. (1951). *Social Choice and Individual Values*. Wiley: New York.

Barocas, S. and A. D. Selbst (2016). Big Data's Disparate Impact. *California Law Review 104*(3), 671–732.

Bathaee, Y. (2018). The Artificial Intelligence Black Box and the Failure of Intent and Causation. *Harvard Journal of Law & Technology 31*, 889.

Bourbaki, N. (1987). *Elements of Mathematics: Topological Vectors Spaces*. Springer-Verlag Berlin.

Caticha, A. (2004). Relative Entropy and Inductive Inference. In G. Erickson and Y. Zhai (Eds.), *Bayesian Inference and Maximum Entropy Methods In Science and Engineering*, Volume 707, pp. 75–96.

Chambers, C. P. and T. Hayashi (2010). Bayesian Consistent Belief Selection. *Journal of*

Economic Theory 145(1), 432–439.

Cobzas, S. (2006). Fixed point theorems in locally convex spaces - The Schauder mapping method. *Fixed Point Theory and Applications 2006*.

Damiano, E. (2006). Choice under Limited Uncertainty. *The B.E. Journal of Theoretical Economics 6*(1), 1–35.

Dominiak, A., M. Kovach, and G. Tserenjigmid (2021). Minimum Distance Belief Updating with General Information. *Working Paper*.

Doshi-Velez, F. and B. Kim (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv: Machine Learning*.

Epstein, L. G. (2006). An Axiomatic Model of Non-Bayesian Updating. *Review of Economic Studies 73*(2), 413–436.

Gajdos, T., T. Hayashi, J.-M. Tallon, and J.-C. Vergnaud (2008). Attitude toward Imprecise Information. *Journal of Economic Theory 140*(1), 27–65.

Gennaioli, N. and A. Shleifer (2010). What Comes to Mind. *Quarterly Journal of Economics 125*(4), 1399–1433.

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi (2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys 51*(5), 1–42.

Kahneman, D., O. Sibony, and C. R. Sunstein (2021). *Noise: A Flaw in Human Judgment*. Little, Brown.

Ke, S., C. Zhao, Z. Wang, and S.-L. Hsieh (2021). Behavioral Neural Networks. *Working Paper*.

Kingston, J. K. (2016). Artificial Intelligence and Legal Liability. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pp. 269–279. Springer.

Kovach, M. (2021). Conservative Updating. *Working Paper*.

Mullainathan, S., J. Schwartzstein, and A. Shleifer (2008). Coarse Thinking and Persuasion. *Quarterly Journal of Economics 123*(2), 577–619.

Nunes, A., B. Reimer, and J. Coughlin (2018, 04). People Must Retain Control of Autonomous Vehicles. *Nature 556*, 169–171.

Ortoleva, P. (2012). Modeling the Change of Paradigm: Non-Bayesian Reactions to Unexpected News. *American Economic Review 102*(6), 2410–2436.

Pedreshi, D., S. Ruggieri, and F. Turini (2008). Discrimination-Aware Data Mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, pp. 560–568. Association for Computing Machinery.

Rabin, M. (2002). Inference by Believers in the Law of Small Numbers. *Quarterly Journal of Economics 117*(3), 775–816.

Rabin, M. and J. L. Schrag (1999). First Impressions Matter: A Model of Confirmatory Bias. *Quarterly Journal of Economics 114*(1), 37–82.

Ribeiro, M. T., S. Singh, and C. Guestrin (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, New York, NY, USA, pp. 1135–1144. Association for Computing Machinery.

Shore, J. E. and R. W. Johnson (1980). Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy. *IEEE Transactions of Information Theory IT-26*, 26–27.

Skilling, J. (1988). The Axioms of Maximum Entropy. *Maximum-Entropy and Bayesian Methods in Science and Engineering 1*, 173–187.

Varadarajan, V. S. (1958). Weak Convergence of Measures on Separable Metric Spaces. *Sankhyā: The Indian Journal of Statistics (1933-1960) 19*(1/2), 15–22.

Wexler, R. (2017). When a Computer Program Keeps You in Jail: How Computers Are Harming Criminal Justice. *New York Times 13*.

Zhao, C. (2020). Representativeness and similarity. *Working Paper*.

Zhao, C. (2021). Pseudo-Bayesian Updating. *Theoretical Economics*, forthcoming.

Appendix

Proof of Theorem 1

Proof. By way of contradiction suppose monotonicity, partial obedience, sensitivity to repetition, and regularity hold.

First, we show that for any $p < \alpha < 1$, there exists $N \in \mathbb{N}$ such that $p_{[\alpha,1]^N} > \alpha$. If $p_{[\alpha,1]} \leq p$, then applying monotonicity inductively yields $p_{[\alpha,1]^n} \leq p_{[\alpha,1]^{n-1}} \leq \cdots \leq p < \alpha$ for any $n > 1$. It follows that $p_{[\alpha,1]^n}$ is a decreasing sequence that converges to some point $p^* < \alpha$. This violates regularity since $[\alpha, 1], [\alpha, 1], \ldots$ is an reinforcing sequence. Hence, $p_{[\alpha,1]} > p$. Applying monotonicity inductively yields $p_{[\alpha,1]^n} \geq p_{[\alpha,1]^{n-1}} \geq \cdots \geq p$. Thus, $p_{[\alpha,1]^n}$ is an increasing sequence that converges to some $p^* \in [0, 1]$. By regularity, $p^* \geq \alpha$. To establish the claim, it suffices to show that $p^* > \alpha$. Suppose $p^* = \alpha$. By partial obedience, there exists $q < \alpha$ such that $q_{[\alpha,1]} \geq \alpha$. Since $p_{[\alpha,1]^n}$ converges to α, there exists $M \in \mathbb{N}$ such that $p_{[\alpha,1]^M} \geq q$. Then monotonicity implies that $p_{[\alpha,1]^{M+1}} \geq q_{[\alpha,1]} \geq \alpha$. Let $M_0 \leq M$ be a positive interger such that $p_{[\alpha,1]^{M_0-1}} < \alpha$ and $p_{[\alpha,1]^{M_0}} \geq \alpha$. Since $p_{[\alpha,1]^n}$ is an increasing sequence that converges to α, we have $p_{[\alpha,1]^n} = \alpha$ for any $n \geq M_0$. Let $r = p_{[\alpha,1]^{M_0-1}}$. Then we have $r < \alpha$ and $r_{[\alpha,1]^n} = \alpha$ for any $n \geq 1$, which violates sensitivity to repetition. Hence, $p^* > \alpha$ and the claim is established.

Now, fix any $p < \alpha < 1$. Let $I := [\alpha, 1]$, $J := [0, \alpha]$, and $N \in \mathbb{N}$ be such that $p_{I^N} > \alpha$. Without loss of generality assume that $q := p_{I^{N-1}} \leq \alpha$. There are two cases: $q_{IJ} \geq q$ or $q_{IJ} < q$.

Suppose $q_{IJ} \geq q$. Applying monotonicity inductively yields $q_{IJI} \geq q_I$, $q_{(IJ)^2} \geq q_{IJ}$, $q_{(IJ)^2I} \geq q_{IJI}, \ldots$ It follows that $q_{(IJ)^nI} \geq q_{(IJ)^{n-1}I} \geq \cdots \geq q_{IJI} \geq q_I > \alpha$ for any $n > 1$. Thus, $q_{(IJ)^nI}$ does not converge to α, which contradicts regularity.

Suppose $q_{IJ} < q$. Applying monotonicity inductively yields $q_{IJI} \leq q_I$, $q_{(IJ)^2} \leq q_{IJ}$, $q_{(IJ)^2I} \leq q_{IJI}, \ldots$ It follows that $q_{(IJ)^n} \leq q_{(IJ)^{n-1}} \leq \cdots \leq q_{IJ} < q \leq \alpha$ for any $n > 1$. Thus, $q_{(IJ)^n}$ does not converge to α, which contradicts regularity. □

Proof of Lemma 1

Proof. Let a_0 be such that $u(a_0, \theta) = 0$ for all $\theta \in \Theta$. Consider any $I \in \mathcal{I}$ defined by continuous and bounded functions $u_1, u_2, \ldots, u_k$. Let L be such that $|u_i(\theta)| < L$ for all i. By richness, there exists $a_i \in \mathcal{A}$ such that $u(a_i, \cdot) = -\frac{u_i}{L}$ for each i, which implies that $I(a_0, \{a_0, \ldots, a_k\}) = I$. □

Proof of Proposition 1

Proof. Let $W(r) := U(a, r) - U(b, r) = \int_\Theta (u(a, \theta) - u(b, \theta)) r(d\theta)$ for any $r \in \Delta(\Theta)$. Clearly W is a linear function; i.e., $W(\alpha r + (1-\alpha) r') = \alpha W(r) + (1-\alpha) W(r')$ for any $r, r' \in \Delta(\Theta)$.

We prove the "if" part first. Suppose $W(p) \leq W(q)$ and $\alpha p + (1-\alpha) r \in I(a, \{a, b\})$ for some $\alpha \in [0, 1]$ and $r \in \Delta(\Theta)$. Then $W(\alpha p + (1-\alpha) r) \geq 0$. Thus,

$$\begin{aligned} W(\alpha q + (1-\alpha) r) &= \alpha W(q) + (1-\alpha) W(r) \\ &\geq \alpha W(p) + (1-\alpha) W(r) \\ &= W(\alpha p + (1-\alpha) r) \geq 0. \end{aligned}$$

Therefore, $\alpha q + (1-\alpha) r \in I(a, \{a, b\})$.

Now we show the "only if" part. Suppose $\alpha p + (1-\alpha) r \in I(a, \{a, b\})$ implies $\alpha q + (1-\alpha) r \in I(a, \{a, b\})$ for any $\alpha \in [0, 1]$ and $r \in \Delta(\Theta)$. By way of contradiction, assume that $W(p) > W(q)$.

Since $I(a, \{a, b\}) \in \mathcal{I}$, its interior is neither $\emptyset$ nor $\Delta(\Theta)$. Hence, there exists $r \in \Delta(\Theta)$ such that $W(r) < 0$. We show that there also exists $r' \in \Delta(\Theta)$ such that $W(r') > 0$. Suppose not. Then there exists $r_0 \in \text{int}(I(a, \{a, b\})$ such that $W(r_0) = 0$. Consider $r_n = \frac{n-1}{n} r_0 + \frac{1}{n} r$. It is clear that r_n converges to r_0 weakly. However, $r_n \notin I(a, \{a, b\})$ for each n, which contradicts $r_0 \in \text{int}(I(a, \{a, b\})$.

Suppose $W(p) = 0$. Then $W(q) < 0$ and we obtain a contradiction to $p \sqsubset_b^a q$. Suppose

$W(p) < 0$; then linearity of W implies that there exists $\alpha \in (0,1)$ such that

$$W(\alpha p + (1-\alpha)r) > 0 > W(\alpha q + (1-\alpha)r),$$

which contradicts $p \sqsubset_b^a q$. Suppose $W(p) > 0$; then linearity of W implies that there exists $\alpha \in (0,1)$ such that

$$W(\alpha p + (1-\alpha)r') > 0 > W(\alpha q + (1-\alpha)r'),$$

which again contradicts $p \sqsubset_b^a q$. Hence, $W(p) \leq W(q)$. □

Proof of Theorem 2 and Proposition 2

Proof. We will only prove Proposition 2, since it implies Theorem 2. To see that, suppose the sequence of recommendations $\{I_n\}_{n=1}^{\infty}$ reveals the equivalence of a, b. For simplicity of exposition let $p_n := p_{I_1 I_2 \ldots I_n}$. Regularity requires that each accumulation point of $\{p_n\}_{n=1}^{\infty}$ be in $I(a,\{a,b\}) \cap I(b,\{a,b\})$. We show that in this case $U(a,p_n) - U(b,p_n)$ converges to 0.

Suppose there exists $\varepsilon > 0$ and a subsequence $\{p_{n_j}\}_{j=1}^{\infty}$ such that $|U(a,p_{n_j}) - U(b,p_{n_j})| \geq \varepsilon$ for any j. Clearly $K := \{p \in \Delta(\Theta) \mid \left|\int_\Theta (u(a,\theta) - u(b,\theta))p(d\theta)\right| \geq \varepsilon\}$ is closed, since $u(a,\cdot) - u(b,\cdot)$ is bounded and continuous. Furthermore, since Θ is compact, $\Delta(\Theta)$ is also compact (see Varadarajan (1958), Theorem 3.4). Hence K is also compact (and sequentially compact, since $\Delta(\Theta)$ with the topology of weak convergence is metrizable). Thus, $\{p_{n_j}\}_{j=1}^{\infty}$ has a subsequence that converges weakly to some point in K. It follows that $\{p_n\}_{n=1}^{\infty}$ has an accumulation point in K, which contradicts regularity since $K \cap I(a,\{a,b\}) \cap I(b,\{a,b\}) = \emptyset$. Thus, Proposition 2 implies Theorem 2.

Now we proceed to prove Proposition 2. We first prove a useful lemma.

Lemma 2. *Suppose weak monotonicity, partial obedience, sensitivity to repetition*, and weak regularity hold. Then for any $a, b \in \mathcal{A}$ such that $I(a,\{a,b\}) \in \mathcal{I}$ and $p \in I(b,\{a,b\})$, there exists some $N \geq 1$ such that $U(a, p_{I(a,\{a,b\})^N}) > U(b, p_{I(a,\{a,b\})^N})$.*

Proof. Let $I := I(a, \{a, b\})$ and $J := I(b, \{a, b\})$. Let $W(r) := U(a, r) - U(b, r) = \int_\Theta (u(a, \theta) - u(b, \theta))r(d\theta)$ for any $r \in \Delta(\Theta)$. Suppose $p \in J$, i.e., $W(p) \leq 0$. By partial obedience, there exists some $q \notin I$ such that $q_I \in I$. In other words, $W(q) < 0$ and $W(q_I) \geq 0$.

We first show that there exists $M \in \mathbb{N}$ such that $W(p_{I^M}) \geq W(q)$. If $W(p) \geq W(q)$, set $M = 0$ and we are done. If $W(p) < W(q)$, then $p \notin I$. By weak regularity (and the sequential compactness of $\Delta(\Theta)$), $\{p_{I^k}\}_{k=1}^\infty$ has a subsequence that converges to some $r \in I$ weakly. Since $u(a, \cdot) - u(b, \cdot)$ is continuous and bounded, $W(p) < W(q) < 0$, and $W(r) \geq 0$, there exists $M \in \mathbb{N}$ such that $W(p_{I^M}) \geq W(q)$.

Now we show that if $W(p) \leq 0$, then there exists $N \geq 1$ such that $W(p_{I^N}) > 0$. Let $M \in \mathbb{N}$ be such that $W(p_{I^M}) \geq W(q)$. Suppose $W(q_I) > 0$. Then by weak monotonicity and Proposition 1, $W(p_{I^M}) \geq W(q)$ implies $W(p_{I^{M+1}}) \geq W(q_I) > 0$ and we are done. Suppose $W(q_I) = 0$. By weak monotonicity and Proposition 1, since $W(q) < 0 = W(q_I)$, we have $W(q) \leq W(q_I) \leq W(q_{I^2}) \leq \dots$. Thus, $W(q_{I^n}) \geq 0$ for any $n \geq 1$. By sensitivity to repetition*, there exists $N_0 \geq 1$ such that $W(q_{I^{N_0}}) > 0$. Applying weak monotonicity and Propsition 1 inductively, $W(p_{I^{M+N_0}}) \geq W(q_{I^{N_0}}) > 0$. □

Now we are ready to prove Proposition 2. Consider any $a, b \in \mathcal{A}$ such that $I(a, \{a, b\})$, $I(b, \{a, b\}) \in \mathcal{I}$. For simplicity of exposition, let $I := I(a, \{a, b\})$, $J := I(b, \{a, b\})$, and define $W(r) := U(a, r) - U(b, r)$ for any $r \in \Delta(\Theta)$. For any $p \in \Delta(\Theta)$, since $I \cup J = \Delta(\Theta)$, we have either $p \in I$ or $p \in J$. We will only prove the case in which $p \in J$, since the other case is symmetric.

By Lemma 2, since $p \in J$ (and thus $W(p) \leq 0$), there exists $N \geq 1$ such that $W(p_{I^N}) > 0$. Consider $p_{I^N J^N}$. There are two cases: $W(p_{I^N J^N}) \geq W(p)$ or $W(p_{I^N J^N}) < W(p)$.

Suppose $W(p_{I^N J^N}) \geq W(p)$. We show that $W(p_{(I^N J^N)^n}) \geq W(p)$ and $W(p_{(I^N J^N)^n I^N}) \geq W(p_{I^N})$ for any $n \geq 1$. By weak monotonicity and Lemma 1, we have $W(p_{I^N J^N I}) \geq W(p_I)$, $W(p_{I^N J^N I^2}) \geq W(p_{I^2}), \dots, W(p_{I^N J^N I^N}) \geq W(p_{I^N})$, and thus the claim holds for $n = 1$. Suppose $W(p_{(I^N J^N)^k}) \geq W(p)$ and $W(p_{(I^N J^N)^k I^N}) \geq W(p_{I^N})$ for some $k \geq 1$.

Since $J \in \mathcal{I}$, by Lemma 1, $q \sqsubset_a^b r$ if and only if $W(q) \geq W(r)$ for any $q, r \in \Delta(\Theta)$. Then, $W(p_{(I^N J^N)^k I^N}) \geq W(p_{I^N})$ and weak monotonicity with respect to J together yield $W(p_{(I^N J^N)^{k+1}}) \geq W(p_{I^N J^N}) \geq W(p)$. Then, applying weak monotonicity with respect to I yields $W(p_{(I^N J^N)^{k+1} I^N}) \geq W(p_{I^N J^N I^N}) \geq W(p_{I^N})$, which establishes the claim. Since $W(p_{(I^N J^N)^n I^N}) \geq W(p_{I^N}) > 0$ for any $n \geq 1$, $W(p_{(I^N J^N)^n I^N})$ does not converge to 0 as $n \to \infty$.

Now suppose $W(p_{I^N J^N}) < W(p)$. Since $W(p) \leq 0$, we have $W(p_{I^N J^N}) < 0$. We show that $W(p_{(I^N J^N)^n}) \leq W(p_{I^N J^N})$ and $W(p_{(I^N J^N)^n I^N}) \leq W(p_{I^N})$ for any $n \geq 1$. It is clear that by weak monotonicity and Lemma 1, $W(p_{I^N J^N I^N}) \leq W(p_{I^N})$, and thus the claim holds for $n = 1$. Suppose $W(p_{(I^N J^N)^k}) \leq W(p_{I^N J^N}) < 0$ and $W(p_{(I^N J^N)^k I^N}) \leq W(p_{I^N})$ for some $k \geq 1$. Since $J \in \mathcal{I}$, by Lemma 1, $q \sqsubset_a^b r$ if and only if $W(r) \leq W(q)$ for any $q, r \in \Delta(\Theta)$. Then, $W(p_{(I^N J^N)^k I^N}) \leq W(p_{I^N})$ and weak monotonicity with respect to J together yield $W(p_{(I^N J^N)^{k+1}}) \leq W(p_{I^N J^N})$. Then, applying weak monotonicity with respect to I yields $W(p_{(I^N J^N)^{k+1} I^N}) \leq W(p_{I^N J^N I^N}) \leq W(p_{I^N})$, which establishes the claim. Since $W(p_{(I^N J^N)^n}) \leq W(p_{I^N J^N}) < 0$ for any $n \geq 1$, $W(p_{(I^N J^N)^n})$ does not converge to 0 as $n \to \infty$.

Hence, in both cases, the conclusion of Proposition 2 holds. □

Proof of Theorem 3

Proof. We will first prove the "only if" part. The proof is broken into a series of lemmas. Let the set of finite signed measures on (Θ, Σ) be Δ^*. It is clear that Δ^* is a real vector space.

The following lemma regarding Δ^* will be useful.

Lemma 3. *For any $p, q \in \Delta^*$, if $p(S) = q(S)$ for any open subset $S \in \Sigma$, then $p = q$.*

Proof. Clearly the collection of all open sets, denoted as $\mathcal{O}$, is a π-system—it is closed under finite intersections. Let $\mathcal{E} := \{S \in \Sigma | p(S) = q(S)\}$. We know that $\mathcal{O} \subseteq \mathcal{E}$. Now we show that $\mathcal{E}$ is a λ-system. First, it is clear that $\Omega \in \mathcal{O} \subseteq \mathcal{E}$. Second, if $S \in \mathcal{E}$, then since $\Omega \in \mathcal{E}$,

and p, q are finite and additive, $\Omega \backslash S \in \mathcal{E}$. By the countable additivity of p, q, if $S_n \in \mathcal{E}$ for each n and $S_i \cap S_j = \emptyset$, then

$$p\left(\bigcup_{n=1}^{\infty} S_n\right) = \sum_{n=1}^{\infty} p(S_n) = \sum_{n=1}^{\infty} q(S_n) = q\left(\bigcup_{n=1}^{\infty} S_n\right).$$

Thus, $\bigcup_{n=1}^{\infty} S_n \in \mathcal{E}$. By Dynkin's π-λ theorem, the σ-algebra generated by $\mathcal{O}$ is a subset of $\mathcal{E}$, i.e., $\Sigma \subseteq \mathcal{E}$. Thus, $p(S) = q(S)$ for any $S \in \Sigma$. □

Lemma 4. *If $(U, \mathcal{A})$ is rich, for any continuous and bounded function $\tilde{u}$, there exist $a, b \in \mathcal{A}$ and $\lambda > 0$ such that $u(a, \cdot) - u(b, \cdot) = \lambda \tilde{u}$.*

Proof. Suppose $|\tilde{u}(\theta)| < L$ for any $\theta \in \Theta$. Let $\lambda = \frac{1}{L}$. Then $|\lambda \tilde{u}| < 1$. Thus, by the richness condition, there exists $a \in \mathcal{A}$ such that $u(a, \cdot) = \lambda \tilde{u}$. Let $b \in \mathcal{A}$ be such that $u(b, \cdot) = 0$ and we are done. □

A function $\tilde{u} : \Theta \to \mathbb{R}$ is *indefinite* if there exist $\theta, \theta' \in \Theta$ such that $\tilde{u}(\theta) > 0 > \tilde{u}(\theta')$.

Lemma 5. *Suppose strong monotonicity holds. Then for any continuous and bounded function $\tilde{u}$, $p, q \in \Delta(\Theta)$, and $I \in \mathcal{I}$, $\int_\Theta \tilde{u} dp \leq \int_\Theta \tilde{u} dq$ implies $\int_\Theta \tilde{u} dp_I \leq \int_\Theta \tilde{u} dq_I$.*

Proof. Let $\tilde{u}$ be a continuous and bounded function. If $\tilde{u}$ is constant, then the condition in the lemma holds trivially. Suppose there exists $\theta, \theta' \in \Theta$ such that $\tilde{u}(\theta) \neq \tilde{u}(\theta')$. Then clearly there exists $\delta \in \mathbb{R}$ such that $\tilde{u} + \delta$ is indefinite. By Lemma 4, there exist $a, b \in \mathcal{A}$ and $\lambda > 0$ such that $u(a, \cdot) - u(b, \cdot) = \lambda(\tilde{u} + \delta)$. Since $\tilde{u} + \delta$ is indefinite, there exist $\theta, \theta' \in \Theta$ such that $u(a, \theta) - u(b, \theta) > 0 > u(a, \theta') - u(b, \theta')$. Clearly, there exists $p_1, p_2 \in \Delta(\Theta)$ such that $U(a, p_1) > U(b, p_1)$ and $U(a, p_2) < U(b, p_2)$ (taking the corresponding Dirac measures will suffice). Hence, $I(a, \{a, b\})$ is nontrivial. By Proposition 1 and strong monotonicity, $U(a, p) - U(b, p) \leq U(a, q) - U(b, q)$ implies $U(a, p_I) - U(b, p_I) \leq U(a, q_I) - U(b, q_I)$. It follows that $\int_\Theta \tilde{u} dp \leq \int_\Theta \tilde{u} dq$ implies $\int_\Theta \tilde{u} dp_I \leq \int_\Theta \tilde{u} dq_I$. □

A function $\hat{u} : \Theta \to \mathbb{R}$ is a *simple function* if there exists $k \in \mathbb{N}$, $S_1, S_2, \ldots, S_k \in \Sigma$, and $\alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{R}$ such that $\hat{u} = \sum_{i=1}^{k} \alpha_i \mathbf{1}_{S_i}$. A function $\hat{u} : \Theta \to \mathbb{R}$ is a *step function* if there exists $k \in \mathbb{N}$, open subsets $S_1, S_2, \ldots, S_k \in \Sigma$, and $\alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{R}$ such that $\hat{u} = \sum_{i=1}^{k} \alpha_i \mathbf{1}_{S_i}$.

Lemma 6. *Suppose strong monotonicity holds. Then for any step function $\hat{u}$ and $p, q \in \Delta(\Theta)$, $\left(\int_\Theta \hat{u} dp - \int_\Theta \hat{u} dq\right)\left(\int_\Theta \hat{u} dp_I - \int_\Theta \hat{u} dq_I\right) \geq 0$.*

Proof. First, note that Θ is compact metric space. Therefore, we may, without loss of generality, assume that $d(\theta, \theta') \leq 1$ for any $\theta, \theta' \in \Theta$. For any $\theta \in \Theta$ and any nonempty subset $S \subseteq \Theta$, let $d(\theta, S) = \inf_{\theta' \in S} d(\theta, \theta')$ and $d(\theta, \emptyset) = 1$. The first step is to show that $d(\cdot, S)$ is continuous for any $S \subseteq \Theta$. If $S = \emptyset$ there is nothing to prove. If $S \neq \emptyset$, for any $\theta_0 \in S$ and $\theta, \theta' \in \Theta$, we have

$$d(\theta, S) \leq d(\theta, \theta_0) \leq d(\theta, \theta') + d(\theta', \theta_0),$$

which implies that $d(\theta, S) \leq d(\theta, \theta') + d(\theta', S)$. It follows that $|d(\theta, S) - d(\theta', S)| \leq d(\theta, \theta')$, and thus, $d(\cdot, S)$ is continuous.

For any open subset $S \subseteq \Theta$, let $S^n := \{\theta \in \Theta | d(\theta, \Theta \backslash S) \geq \frac{1}{n}\}$. It is clear that $S^n \subseteq S^{n+1}$ for all n, and $S = \bigcup_{n=1}^{\infty} S^n$. Define $u_n : \Theta \to \mathbb{R}$ as follows:

$$u_n(\theta) = \frac{d(\theta, \Theta \backslash S)}{d(\theta, \Theta \backslash S) + d(\theta, S^n)}.$$

Clearly u_n is continuous and bounded. In particular, $u_n(\theta) = 0$ if $\theta \notin S$; $u_n(\theta) = 1$ if $\theta \in S^n$; $u_n(\theta) \in [0, 1]$ if $\theta \in S \backslash S^n$. Furthermore, u_n converges pointwise to $\mathbf{1}_S$ and $|u_n(\theta)| \leq \mathbf{1}_S(\theta)$ for any $\theta \in \Theta$.

Consider any step function $\hat{u} = \sum_{i=1}^{k} \alpha_i \mathbf{1}_{S_i}$. Approximate each $\mathbf{1}_{S_i}$ with $u_{i,n}$ as above.

For each n, $\sum_{i=1}^{k} \alpha_i u_{i,n}$ is continuous and bounded. Thus, by Lemma 5, for each n,

$$\left(\int_\Theta \sum_{i=1}^{k} \alpha_i u_{i,n} dp - \int_\Theta \sum_{i=1}^{k} \alpha_i u_{i,n} dq\right)\left(\int_\Theta \sum_{i=1}^{k} \alpha_i u_{i,n} dp_I - \int_\Theta \sum_{i=1}^{k} \alpha_i u_{i,n} dq_I\right) \geq 0.$$

Then Lebesgue's dominated convergence theorem completes the proof. □

Lemma 7. *Suppose strong monotonicity holds. Then for any $p, q \in \Delta(\Theta) \subseteq \Delta^*$, and $I \in \mathcal{I}$, there exists $\varepsilon \geq 0$ such that $p_I - q_I = \varepsilon(p - q)$.*

Proof. If $p = q$ there is nothing to prove. Suppose $p \neq q$. Then by Lemma 3 there exists an open subset S such that $p(S) \neq q(S)$. Let $\varepsilon := \frac{p_I(S) - q_I(S)}{p(S) - q(S)}$. Again by Lemma 3, to show that $p_I - q_I = \varepsilon(p - q)$, since $p_I - q_I$ and $\varepsilon(p - q)$ are both elements of Δ^*, it suffices to show that $p_I(T) - q_I(T) = \varepsilon(p(T) - q(T))$ for any open subset $T \in \Sigma$.

Consider any step function of the form $\hat{u} = \alpha \mathbf{1}_S + \mathbf{1}_T$. By Lemma 6, for any $\alpha \in \mathbb{R}$, we have

$$(\alpha(p(S) - q(S)) + p(T) - q(T))(\alpha(p_I(S) - q_I(S)) + p_I(T) - q_I(T)) \geq 0. \tag{1}$$

Suppose for some $\alpha \in \mathbb{R}$, $\alpha(p(S) - q(S)) + p(T) - q(T) = 0$ but $\alpha(p_I(S) - q_I(S)) + p_I(T) - q_I(T) \neq 0$. Then, since $p(S) \neq q(S)$, there always exists α' which is close to α such that

$$(\alpha'(p(S) - q(S)) + p(T) - q(T))(\alpha'(p_I(S) - q_I(S)) + p_I(T) - q_I(T)) < 0,$$

which is a contraction. Hence, for any $\alpha \in \mathbb{R}$, if $\alpha(p(S) - q(S)) + p(T) - q(T) = 0$, then $\alpha(p_I(S) - q_I(S)) + p_I(T) - q_I(T) = 0$.

Since $p(S) \neq q(S)$, $\alpha(p(S) - q(S)) + p(T) - q(T) = 0$ if and only if

$$\alpha = -\frac{p(T) - q(T)}{p(S) - q(S)}.$$

Thus, it must be the case that

$$-(p(T)-q(T))\frac{p_I(S)-q_I(S)}{p(S)-q(S)}+p_I(T)-q_I(T)=0$$

which implies

$$p_I(T)-q_I(T)=\varepsilon(p(T)-q(T)). \tag{2}$$

Furthermore, equations (1) and (2) imply that

$$\varepsilon(\alpha(p(S)-q(S))+p(T)-q(T))^2\geq 0$$

for any $\alpha\in\mathbb{R}$. Since $p(S)\neq q(S)$, it is clear that $\varepsilon\geq 0$. □

Lemma 8. *Suppose strong monotonicity holds and* $|\Theta|\geq 3$. *Then for any* $I\in\mathcal{I}$, *there exists* $\varepsilon\geq 0$ *such that* $p_I-q_I=\varepsilon(p-q)$ *for any* $p,q\in\Delta(\Theta)\subseteq\Delta^*$.

Proof. Consider any $p,q,r\in\Delta(\Theta)\subseteq\Delta^*$ that are linearly independent. Since $|\Theta|\geq 3$, such p,q,r exist. By the previous lemma, suppose that $p_I-q_I=\varepsilon_1(p-q)$, $q_I-r_I=\varepsilon_2(q-r)$, and $r_I-p_I=\varepsilon_3(r-p)$. It follows that

$$0=\varepsilon_1(p-q)+\varepsilon_2(q-r)+\varepsilon_3(r-p),$$

which, since p,q,r are linearly independent, implies that $\varepsilon_1=\varepsilon_2=\varepsilon_3:=\varepsilon$.

Now let $p',q'\in\Delta(\Theta)\subseteq\Delta^*$ with $p'\neq q'$. By the previous lemma, there exists $\varepsilon'\geq 0$ such that $p'_I-q'_I=\varepsilon'(p'-q')$. We now show that $\varepsilon'=\varepsilon$.

First, we show that there must exist $\tilde{p}\in\{p,q,r\}$ such that $\tilde{p},p',q'$ are linearly independent. By way of contradiction, suppose that $\tilde{p},p',q'$ are linearly dependent for each $\tilde{p}\in\{p,q,r\}$. In other words, for each $\tilde{p}\in\{p,q,r\}$, there exists $\alpha,\beta,\gamma\in\mathbb{R}$ such that at least

one of them is nonzero, and that

$$\alpha\tilde{p} + \beta p' + \gamma q' = 0.$$

Note that $p', q' \in \Delta(\Theta)$ and $p' \neq q'$ together imply that p', q' are linearly independent. Since p', q' are linearly independent, $\alpha \neq 0$. Thus, for each $\tilde{p} \in \{p, q, r\}$, there exists $\tilde{\alpha}, \tilde{\beta} \in \mathbb{R}$ such that $\tilde{p} = \tilde{\alpha}p' + \tilde{\beta}q'$, which contradicts the fact that p, q, r are linearly independent.

Without loss of generality, assume that $\tilde{p} = p$, and thus p, p', q' are linearly independent. By the same argument as in the first paragraph, it follows that $p_I - p'_I = \varepsilon'(p - p')$. The last step is to show that there must exist $\tilde{q} \in \{q, r\}$ such that $p, \tilde{q}, p'$ are linearly independent. Suppose not. Since p, p' are linearly independent, there exist $\alpha_1, \alpha_2, \beta_1, \beta_2 \in \mathbb{R}$ such that

$$q = \alpha_1 p + \beta_1 p'$$
$$r = \alpha_2 p + \beta_2 p',$$

which contradicts the fact that p, q, r are linearly dependent.

Without loss of generality, assume that $\tilde{q} = q$, and thus p, q, p' are linearly independent. Again by the same argument as in the first paragraph, $p_I - q_I = \varepsilon'(p - q)$. Then $p \neq q$ implies that $\varepsilon = \varepsilon'$, which establishes the lemma. □

Lemma 9. *Suppose strong monotonicity holds and $|\Theta| \geq 3$. Then for any $I \in \mathcal{I}$, p_n converges to p weakly implies that $(p_n)_I$ converges to p_I weakly.*

Proof. By the previous lemma we know that there exists $\varepsilon \geq 0$ such that, for any n,

$$(p_n)_I - p_I = \varepsilon(p_n - p).$$

To establish the lemma, it suffices to show that for any bounded continuous function f,

$$\int_\Theta f d(p_n)_I - \int_\Theta f dp_I = \varepsilon \left(\int_\Theta f dp_n - \int_\Theta f dp \right). \tag{3}$$

By the definition of Lebesgue integral, we only need to show (3) if f is nonnegative, bounded, and continuous.

Consider any simple function

$$g = \sum_{i=1}^{n} \alpha_i \mathbf{1}_{S_i}$$

in which $\alpha_i \geq 0$, $S_i \in \Sigma$, and $S_i \cap S_j = \emptyset$ for all i, j. It is clear that

$$\begin{aligned} \int_\Theta g d(p_n)_I - \int_\Theta g dp_I &= \sum_{i=1}^{n} \alpha_i [(p_n)_I(S_i) - p_I(S_i)] \\ &= \varepsilon \sum_{i=1}^{n} \alpha_i [p_n(S_i) - p(S_i)] \\ &= \varepsilon \left(\int_\Theta g dp_n - \int_\Theta g dp \right). \end{aligned}$$

Let g_k be a sequence of simple functions such that g_k converges pointwise to f, and $0 \leq g_k(\theta) \leq g_{k+1}(\theta)$ for all k and $\theta \in \Theta$. Then

$$\int_\Theta g_k d(p_n)_I - \int_\Theta g_k dp_I = \varepsilon \left(\int_\Theta g_k dp_n - \int_\Theta g_k dp \right).$$

By Beppo Levi's monotone convergence theorem, letting $k \to \infty$, we obtain

$$\int_\Theta f d(p_n)_I - \int_\Theta f dp_I = \varepsilon \left(\int_\Theta f dp_n - \int_\Theta f dp \right)$$

and the lemma is established. □

Lemma 10. *Suppose strong monotonicity holds and $|\Theta| \geq 3$. Then for any $I \in \mathcal{I}$, there exists $q \in \Delta(\Theta)$ such that $q_I = q$. Thus, for any $I \in \mathcal{I}$, there exists $q \in \Delta(\Theta)$ and $\varepsilon \in [0, 1]$*

such that

$$p_I = \varepsilon p + (1 - \varepsilon)q$$

for any $p \in \Delta(\Theta)$.

Proof. Let $C(\Theta)$ be the set of continuous real-valued functions defined on Θ. Since Θ is compact, each $f \in C(\Theta)$ is also bounded and uniformly contiuous. We equip Δ^* with the weak topology $\sigma(\Delta^*, C(\Theta))$, i.e., the topology of weak convergence. It is well known that (i) $\Delta(\Theta)$ is compact (see Varadarajan (1958), Theorem 3.4); (ii) $\sigma(\Delta^*, C(\Theta))$ is locally convex (see Bourbaki (1987), page II.40-42); (iii) $\sigma(\Delta^*, C(\Theta))$ is Hausdorff (see Bourbaki (1987), page II.41, Proposition 1; page II.43, Proposition 2; and Varadarajan (1958), Lemma 2.3).

Thus, Δ^* is a Hausdorff locally convex topological vector space, and $\Delta(\Theta)$ is a convex and compact subset of Δ^*. By the previous lemma, for each $I \in \mathcal{I}$, the mapping $\pi(\cdot, I) : \Delta(\Theta) \to \Delta(\Theta)$ is continuous. By the Schauder–Tychonoff fixed point theorem (see Cobzas (2006), Theorem 2.3 for the exact version of the theorem and a proof), there exists $q \in \Delta(\Theta)$ such that $q_I = q$. Then by Lemma 8, for any $I \in \mathcal{I}$, there exists $q \in \Delta(\Theta)$ and $\varepsilon \geq 0$ such that

$$p_I = \varepsilon p + (1 - \varepsilon)q$$

for any $p \in \Delta(\Theta)$. It follows that

$$p_{I^n} = \varepsilon^n p + (1 - \varepsilon^n)q$$

for any $n \in \mathbb{N}$ and $p \in \Delta(\Theta)$.

Suppose $\varepsilon > 1$. Pick any $p \neq q$. Then there exists $S \in \Sigma$ such that $p(S) - q(S) < 0$. It follows that there exists n large enough such that $p_{I^n}(S) = q(S) + \varepsilon^n(p(S) - q(S)) < 0$, which contradicts the definition of an updating rule. Hence, we conclude that $\varepsilon \in [0, 1]$. □

Lemma 11. *Suppose strong monotonicity, weak regularity, and sensitivity to repetition**

hold, and $|\Theta| \geq 3$. *Then for any* $I \in \mathcal{I}$, *there exists* $q \in I$ *and* $\varepsilon \in (0,1)$ *such that*

$$p_I = \varepsilon p + (1-\varepsilon)q$$

for any $p \in \Delta(\Theta)$.

Proof. Fix $I \in \mathcal{I}$. By the previous lemma, there exists $q \in \Delta(\Theta)$ such that $q_I = q$. If $q \notin I$, then the sequence q_{I^n} does not have any accumulation point in I, which contradicts weak regularity. Hence, any fixed point of the mapping $\pi(\cdot, I)$ must be in I. Hence, there exist $q \in I$ and $\varepsilon \in [0,1]$ such that

$$p_I = \varepsilon p + (1-\varepsilon)q$$

for any $p \in \Delta(\Theta)$.

Let $I = I(a_1, \{a_1, a_2, \ldots, a_k\})$. By richness, such $a_1, a_2, \ldots, a_k$ exist for any $I \in \mathcal{I}$. Since $I \neq \Delta(\Theta)$, there exist $1 < i \leq k$ and $p \in \Delta(\Theta)$ such that $p \notin I(a_1, \{a_1, a_i\})$. In addition, it is clear that $I \subseteq I(a_1, \{a_1, a_i\})$. Since I is nontrivial and $p \notin I(a_1, \{a_1, a_i\})$, $I(a_1, \{a_1, a_i\})$ is also nontrivial. By sensitivity to repetition* and Proposition 1, there exists $m, n \geq 1$ such that $m \neq n$ and $U(a_1, p_{I^m}) - U(a_i, p_{I^m}) > U(a_1, p_{I^n}) - U(a_i, p_{I^n})$. It follows that $\varepsilon \in (0,1)$, and we are done. □

Now we show the "if" part. It is easy to see that any contraction rule will satisfy weak regularity.

To show strong monotonicity, we first show that for any continuous and bounded function f and $I \in \mathcal{I}$,

$$\int_\Theta f dp \geq \int_\Theta f dq \quad \Rightarrow \quad \int_\Theta f dp_I \geq \int_\Theta f dq_I. \tag{4}$$

Note that any contraction rule satisfies

$$p_I - q_I = \varepsilon(I)(p - q)$$

for any $p, q \in \Delta(\Theta)$. Using the same argument as the one for (3), we have

$$\int_\Theta f dp_I - \int_\Theta f dq_I = \varepsilon(I)\left(\int_\Theta f dp - \int_\Theta f dq\right),$$

which implies (4).

Now suppose $p \sqsubset_b^a q$. If $I(a, \{a, b\}) \in \mathcal{I}$, then by Proposition 1 and equation (4) we are done. If $\text{int}(I(a, \{a, b\})) = \Delta(\Theta)$, then $I(a, \{a, b\}) = \Delta(\Theta)$, and thus $r \sqsubset_b^a s$ for any $r, s \in \Delta(\Theta)$ and we are done. If $\text{int}(I(a, \{a, b\})) = \emptyset$, then $I(b, \{a, b\}) = \Delta(\Theta)$, i.e., $U(a, s) - U(b, s) \leq 0$ for any $s \in \Delta(\Theta)$. Thus, $s \in I(a, \{a, b\})$ implies $U(a, s) - U(b, s) = 0$. Consider $r \in \Delta(\Theta)$ and $\alpha \in [0, 1]$ such that $\alpha p_I + (1 - \alpha)r \in I(a, \{a, b\})$. If $p_I \notin I(a, \{a, b\})$, then $\alpha = 0$ and $r \in I(a, \{a, b\})$. Therefore, $\alpha q_I + (1 - \alpha)r = r \in I(a, \{a, b\})$ and we are done. If $p_I \in I(a, \{a, b\})$, then $U(a, p_I) - U(b, p_I) = 0$. Since $\varepsilon(I) \in (0, 1)$, we must have $U(a, p) - U(b, p) = U(a, \rho(I)) - U(b, \rho(I)) = 0$. Since $p \in I(a, \{a, b\})$ and $p \sqsubset_b^a q$, we have $\frac{1}{2}q + \frac{1}{2}p \in I(a, \{a, b\})$, which implies that $q \in I(a, \{a, b\})$. Therefore, $q_I \in I(a, \{a, b\})$ and $U(a, q_I) - U(b, q_I) = 0 = U(a, p_I) - U(b, p_I)$. Then it is clear that $\alpha q_I + (1-\alpha)r \in I(a, \{a, b\})$.

To show sensitivity to repetition*, let $a, b \in \mathcal{A}$, $p \notin I(a, \{a, b\})$, and $I \in \mathcal{I}$ with $I \subseteq I(a, \{a, b\})$. Since $p \notin I(a, \{a, b\})$ and $\rho(I) \in I \subseteq I(a, \{a, b\})$, we have $U(a, p) - U(b, p) < 0$ and $U(a, \rho(I)) - U(b, \rho(I)) \geq 0$. Thus

$$U(a, p_{I^n}) - U(b, p_{I^n}) = \varepsilon^n(U(a, p) - U(b, p)) + (1 - \varepsilon^n)(U(a, \rho(I)) - U(b, \rho(I))).$$

Since $\varepsilon \in (0, 1)$, it is clear that there exists $m, n \geq 1$ such that $U(a, p_{I^m}) - U(b, p_{I^m}) > U(a, p_{I^n}) - U(b, p_{I^n})$. Note that since $p \notin I(a, \{a, b\})$ and $I \subseteq I(a, \{a, b\})$, $I(a, \{a, b\})$ is nontrivial. Hence, by Proposition 1, $p_{I^m} \not\sqsubset_b^a p_{I^n}$. Thus, sensitivity to repetition* holds.

Next, we show that if $\rho(I) \in \text{int}(I)$ for any $I \in \mathcal{I}$, then partial obedience is satisfied. Clearly, under the topology of weak convergence, each $I \in \mathcal{I}$ is closed and has nonempty interior. For any $p \notin I$, we know that p_{I^n} converges weakly to $\rho(I) \in \text{int}(I)$. Then there exists $m \in \mathbb{N}$ such that $p_{I^m} \notin I$ and $(p_{I_m})_I = p_{I^{m+1}} \in I$, establishing partial obedience. □

PROOF OF THEOREM 4

Proof. We first prove a useful lemma. We say that the updating rule satisfies *weak regularity*† if for any $p \in [0,1]$, each accumulation point of $\{p_{\langle I^{M-1}\rangle I^{n-M+1}}\}_{n=M}^{\infty}$ is in I.

Lemma 12. *Suppose $\alpha \in (0,1)$, and monotonicity*†*, partial obedience*†*, sensitivity to repetition*†*, and weak regularity*† *hold. For any $p \leq \alpha$, there exists some $n \in \mathbb{N}$ such that $p_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]^n} > \alpha$. For any $p \geq \alpha$, there exists some $n \in \mathbb{N}$ such that $p_{\langle [0,\alpha]^{M-1}\rangle [0,\alpha]^n} < \alpha$.*

Proof. We only prove the first statement with $p \leq \alpha$, since the other one is symmetric. By partial obedience†, there exists some $\beta < \alpha$ such that $\beta_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]} \geq \alpha$. Suppose $\beta_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]} > \alpha$. By weak regularity†, for some sufficiently large $N \in \mathbb{N}$, $p_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]^N} > \beta$. Therefore, by monotonicity†, $p_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]^{N+1}} \geq \beta_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]} > \alpha$. Note that monotonicity† is applicable, since $p_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]^{N+1}}$ and $\beta_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]}$ share the same history of $M-1$ recommendations before updating on $[\alpha,1]$.

Suppose $\beta_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]} = \alpha$. By definition, $\beta_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]^2} = \alpha_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]}$. If $\alpha_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]} \leq \alpha$, by applying monotonicity† inductively, we have $\alpha_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]} \geq \alpha_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]^2} \geq \dots$. It then follows from weak regularity† that $\alpha_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]^n} = \alpha$ for any n, which violates sensitivity to repetition†.

Therefore, we must have $\alpha_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]} > \alpha$. By weak regularity†, for some sufficiently large $N \in \mathbb{N}$, $p_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]^N} > \beta$. Therefore, by monotonicity†, $p_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]^{N+1}} \geq \beta_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]}$. Again by monotonicity†, $p_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]^{N+2}} \geq \beta_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]^2} = \alpha_{\langle [\alpha,1]^{M-1}\rangle [\alpha,1]} > \alpha$. □

Now we are ready to show the impossibility result. Suppose that monotonicity†, partial obedience†, sensitivity to repetition†, and regularity† hold. Fix $\alpha \in (0,1)$ and $p \leq \alpha$. Let $I := [\alpha,1]$, $J := [0,\alpha]$, and $N \geq M$ be such that $p_{\langle I^{M-1}\rangle I^{N-M+1}} > \alpha$. Consider $p_{\langle I^{M-1}\rangle I^{N-M+1} J^N I^{M-1}}$. We have either $p_{\langle I^{M-1}\rangle I^{N-M+1} J^N I^{M-1}} \geq p$ or $p_{\langle I^{M-1}\rangle I^{N-M+1} J^N I^{M-1}} < p$.

Suppose $p_{\langle I^{M-1}\rangle I^{N-M+1} J^N I^{M-1}} \geq p$. We show that $p_{\langle I^{M-1}\rangle I^{N-M+1} J^N (I^N J^N)^{n-1} I^{M-1}} \geq p$ and $p_{\langle I^{M-1}\rangle I^{N-M+1} J^N (I^N J^N)^{n-1} I^N} \geq p_{\langle I^{M-1}\rangle I^{N-M+1}}$ for any $n \geq 1$. By $p_{\langle I^{M-1}\rangle I^{N-M+1} J^N I^{M-1}} \geq$

p and monotonicity†, we have $p_{\langle I^{M-1}\rangle I^{N-M+1}JNI^{M-1}I} \geq p_{\langle I^{M-1}\rangle I}$, $p_{\langle I^{M-1}\rangle I^{N-M+1}JNI^{M-1}I^2} \geq p_{\langle I^{M-1}\rangle I^2}, \ldots, p_{\langle I^{M-1}\rangle I^{N-M+1}JNI^N} \geq p_{\langle I^{M-1}\rangle I^{N-M+1}}$. Note that monotonicity† is applicable in each step since the decision maker's memory is always full of I's in both sides of each inequality. Hence the claim holds for $n = 1$. Suppose that $p_{\langle I^{M-1}\rangle I^{N-M+1}JN(I^NJ^N)^{k-1}I^{M-1}} \geq p$ and $p_{\langle I^{M-1}\rangle I^{N-M+1}JN(I^NJ^N)^{k-1}I^N} \geq p_{\langle I^{M-1}\rangle I^{N-M+1}}$ for $k \geq 1$. Note that $p_{\langle I^{M-1}\rangle I^{N-M+1}JN(I^NJ^N)^{k-1}I^N}$ and $p_{\langle I^{M-1}\rangle I^{N-M+1}}$ share the same memory of $M-1$ I's before learning the last I. Then, if we let the decision maker learns J N times and then I N times, in each step, monotonicity† is applicable. Thus, it follows from $p_{\langle I^{M-1}\rangle I^{N-M+1}JN(I^NJ^N)^{k-1}I^N} \geq p_{\langle I^{M-1}\rangle I^{N-M+1}}$ and monotonicity† that $p_{\langle I^{M-1}\rangle I^{N-M+1}JN(I^NJ^N)^kI^{M-1}} \geq p_{\langle I^{M-1}\rangle I^{N-M+1}JNI^{M-1}} \geq p$ and $p_{\langle I^{M-1}\rangle I^{N-M+1}JN(I^NJ^N)^kI^N} \geq p_{\langle I^{M-1}\rangle I^{N-M+1}JNI^N} \geq p_{\langle I^{M-1}\rangle I^{N-M+1}}$, which establishes the claim. Note that regularity† is violated, since $p_{\langle I^{M-1}\rangle I^{N-M+1}JN(I^NJ^N)^{n-1}I^N} \geq p_{\langle I^{M-1}\rangle I^{N-M+1}} > \alpha$ for any $n \geq 1$.

Suppose $p_{\langle I^{M-1}\rangle I^{N-M+1}JNI^{M-1}} < p$. Since $p \leq \alpha$, $p_{\langle I^{M-1}\rangle I^{N-M+1}JNI^{M-1}} < \alpha$. We show that $p_{\langle I^{M-1}\rangle I^{N-M+1}JN(I^NJ^N)^{n-1}I^{M-1}} \leq p_{\langle I^{M-1}\rangle I^{N-M+1}JNI^{M-1}}$ and $p_{\langle I^{M-1}\rangle I^{N-M+1}JN(I^NJ^N)^{n-1}I^N} \leq p_{\langle I^{M-1}\rangle I^{N-M+1}}$ for any $n \geq 1$. Similarly to the previous case, by $p_{\langle I^{M-1}\rangle I^{N-M+1}JNI^{M-1}} < p$ and monotonicity†, we have $p_{\langle I^{M-1}\rangle I^{N-M+1}JNI^N} \leq p_{\langle I^{M-1}\rangle I^{N-M+1}}$. Hence the claim holds for $n = 1$. Suppose that the claim holds for $n = k \geq 1$, i.e. $p_{\langle I^{M-1}\rangle I^{N-M+1}JN(I^NJ^N)^{k-1}I^{M-1}} \leq p_{\langle I^{M-1}\rangle I^{N-M+1}JNI^{M-1}}$ and $p_{\langle I^{M-1}\rangle I^{N-M+1}JN(I^NJ^N)^{k-1}I^N} \leq p_{\langle I^{M-1}\rangle I^{N-M+1}}$. Similarly to the previous case, it follows from $p_{\langle I^{M-1}\rangle I^{N-M+1}JN(I^NJ^N)^{k-1}I^N} \leq p_{\langle I^{M-1}\rangle I^{N-M+1}}$ and monotonicity† that $p_{\langle I^{M-1}\rangle I^{N-M+1}JN(I^NJ^N)^kI^{M-1}} \leq p_{\langle I^{M-1}\rangle I^{N-M+1}JNI^{M-1}}$, and that $p_{\langle I^{M-1}\rangle I^{N-M+1}JN(I^NJ^N)^kI^N} \leq p_{\langle I^{M-1}\rangle I^{N-M+1}JNI^N} \leq p_{\langle I^{M-1}\rangle I^{N-M+1}}$, which establishes the claim. Note that regularity† is violated, since $p_{\langle I^{M-1}\rangle I^{N-M+1}JN(I^NJ^N)^{n-1}I^{M-1}} \leq p_{\langle I^{M-1}\rangle I^{N-M+1}JNI^{M-1}} < \alpha$ for any $n \geq 1$. $\square$