Algorithmic Assistance with Recommendation-Dependent Preferences

Bryce McLaughlin Jann Spiess
Stanford University

This version: August 15, 2022

Comments welcome!

Abstract

When we use algorithms to produce recommendations, we typically think of these recommendations as providing helpful information, such as when risk assessments are presented to judges or doctors. But when a decision-maker obtains a recommendation, they may not only react to the information. The decision-maker may view the recommendation as a default action, making it costly for them to deviate, for example when a judge is reluctant to overrule a high-risk assessment of a defendant or a doctor fears the consequences of deviating from recommended procedures. In this article, we consider the effect and design of recommendations when they affect choices not just by shifting beliefs, but also by altering preferences. We motivate our model from institutional factors, such as a desire to avoid audits, as well as from well-established models in behavioral science that predict loss aversion relative to a reference point, which here is set by the algorithm. We show that recommendation-dependent preferences create inefficiencies where the decision-maker is overly responsive to the recommendation, which changes the optimal design of the algorithm towards providing less conservative recommendations. As a potential remedy, we discuss an algorithm that strategically withholds recommendations, and show how it can improve the quality of final decisions.

Bryce McLaughlin (brycem@stanford.edu) and Jann Spiess (jspiess@stanford.edu), Graduate School of Business, Stanford University. We thank Talia Gillis and audience members at the 2022 MSOM Conference for helpful comments and suggestions.

1 Introduction

One important application of algorithms is to turn complex data into simple recommendations that help decision-makers take better decisions, such as risk assessments presented to judges or doctors. We typically think of such recommendations as providing additional information about which choices will lead to better outcomes. But decision-makers may react to recommendations not just by shifting beliefs, but also by changing their preferences. In this article, we consider the effect and design of recommendations when they also impose a cost on the decision-maker whenever they deviate from the recommended action, such as when a judge is reluctant to overrule a jailing recommendation or a doctor fears the consequences of deviating from recommended procedure. We show that recommendation dependence creates inefficiencies where the decision-maker is overly responsive to the recommendation, and propose changes to the design of recommendation algorithms towards providing less conservative recommendations.

We consider a decision-maker who takes a binary decision based on their private information along with a binary recommendation provided by an algorithm based on additional data. The decision-maker takes a decision between a safe and a risky action after observing the algorithm's recommendation. When the state of the world is good, the risky action is best, while in the bad state the risky action leads to high loss. The decision-maker therefore uses the available information to form a forecast of the probability that the state is bad, and chooses the risky action only if that predicted probability is low. For example, a judge who considers whether to release a defendant on bail (risky action) aims to release only those defendants with low probability of failing to appear or committing a new crime (bad state).

To this model we add the assumption that recommendations affect decisions not only through the information they provide, but also by setting a reference point for the decision-maker. We assume that the decision-maker perceives an additional (private) cost from making an error when deviating from this recommendation. Specifically, in our model there is an additional loss when the decision-maker takes a risky decision against the safe recommendation of an algorithm when the bad state materializes. Similarly, there may also be an additional loss from deviation when the decision-maker opts for the safe option relative to a risky recommendation when the good state occurs.

A first motivation for such recommendation-dependent preferences stems from institutional factors, such as when deviating from recommendations triggers audits or may create backlash. For example, a judge may be reluctant to release a defendant in light of a jailing recommendation for fear of repercussions, even if they believe that the defendant represents a lower risk. Similarly, a doctor may prefer to order a test (safe decision) when the algorithm recommends so for fear of missing a bad diagnosis against algorithmic advice, and may feel more justified in taking a risky decision when the algorithm concurs.

A second motivation is provided by established models from behavioral science that suggest expected losses impact decision-makers more than commensurate gains, relative to some reference point that we assume here is affected by the algorithmic recommendation. The combination of perceiving decision utility relative to a reference point and experiencing loss aversion relative to that reference point are two of the main features of Prospect Theory, which has been one of the most established frameworks rooted in psychology for describing systematic deviations from rational utility maximization. Here, we show that the combination of those two features predicts recommendation dependence when we assume that the reference point is obtained from the recommended action.

Having set up a model of recommendation-dependent preferences, we show that the effect of algorithmic advice generally differs from a reference-independent baseline case. Recommendation dependence increases adherence to the algorithmic recommendation. This adherence makes decisions less efficient as it reduces the amount of private information that the decision-maker reveals through their chosen action. For example, if a judge is worried about repercussions from releasing a defendant the algorithm recommends to jail, the judge may follow the recommendation even if they have private information that suggests that the defendant is not at high risk of committing a new crime or failing to appear.

Recommendation dependence leads to inefficiencies that can be mitigated (but not completely avoided) by a better design of the recommendation. Under regularity assumptions, we show that if deviating from one of the recommendations becomes more costly then this recommendation should be given less frequently. Specifically, if a decision-maker is reluctant to over-rule a safe recommendation because of additional costs from making a mistake in this case, the algorithm should recommend the safe option less, and instead propose the risky option in some cases where a baseline algorithm would recommend the safe option.

Having shown how recommendation dependence affects the consequences and optimal design of recommendations, we discuss the gain of allowing the recommendation algorithm to strategically withhold information. With recommendation-dependent preferences, adding a third option of not providing a recommendation at all has two distinct benefits. The first benefit is that it allows the transmission of additional information through the recommendation, signaling an intermediate probability of the bad outcome occurring. The second benefit is that not providing a recommendation in some cases also reduces the cost of recommendation dependence, and allows the decision-maker to take optimal decisions in this case. Specifically, we show that adding such an additional "don't know" level within our model improves decisions relatively more in a world with recommendation-dependent preferences than in a world where decision-makers' preferences are not affected by the algorithm.

As an extension, we also consider the case where the decision-maker and the algorithm designer have different preferences, in which case reference dependence turns into a feature that allows the designer to improve outcomes when recommendations are properly designed. This case captures a setting where the decision-maker takes overly risky decision without algorithmic advice, and recommendation-dependent preferences imply that proposing the safe option more often may address this misalignment in risk preferences.

We contribute to cross-disciplinary theoretical and empirical studies on human–computer interaction and critical uses of machine learning, including work where the knowledge of an AI and human decision-makers (or more generally multiple knowledge sources) are combined (e.g. Lawrence et al., 2006; Palley and Soll, 2019), humans assist an AI (e.g. Hampshire et al., 2020; Ibrahim et al., 2021), or an algorithm optimizes advice given to human decision-makers (Bastani et al., 2021). Additional applications and empirical evaluations that motivate our work come from the use of risk assessment tools for pretrial release and sentencing decisions (Stevenson and Doleac, 2019; Imai et al., 2020) and algorithms used by clinicians (Kiani et al., 2020; Murray et al., 2020). Fogliato et al. (2022b) study human overrides of algorithmic recommendations, and argue in favor of human discretion in critical applications. Closely related to our work, Shashikumar et al. (2021) propose adding an "I don't know" (non-)recommendation to improve combined decisions. Raghu et al. (2019) consider the optimization problem of delegating individual instances to an algorithm or a human decision-maker, while (Athey et al., 2020) discuss general trade-offs in the allocation of decision authority between human and AI. Fairness aspects of human–AI systems are considered by e.g. Green and Chen (2019a); Gillis et al. (2021); Donahue et al. (2022); Lima et al. (2022).

We build upon a literature that brings models from psychology into economics and operations. We motivate our model by theories of loss aversion with reference dependence in behavioral economics, going back to Prospect Theory (Kahneman and Tversky, 1979). We contribute to a behavioral operations management literature (recently reviewed by Donohue et al., 2020) that apply behavioral-science theories in operations problems such as the newsvendor and capacity reservation tasks (e.g. Schweitzer and Cachon, 2000; Su, 2008; Johnsen et al., 2019). More specifically, we relate to work that focuses on behavioral aspects in the interaction between human and machine, including algorithm aversion (Dietvorst et al., 2015, 2018), algorithm appreciation (Logg et al., 2019; Bai et al., 2021), and over-reliance on algorithms (Banker and Khetani, 2019; Buçinca et al., 2021). Recent work emphasizes that the success of human–machine collaboration is dependent on details of context, implementation, and presentation of algorithmic advice (such as Bansal et al., 2019a,b; Green and Chen, 2019b; Snyder et al., 2022; Fogliato et al., 2022a), including information about its uncertainty (McGrath et al., 2020; Taudien et al., 2022) and explanations of black-box classifiers (Lakkaraju and Bastani, 2020). Closely related to our work, Fügener et al. (2021) show how over-adherence to AI assistance may reduce diversity of opinions and can lead to worse group decisions. Other recent work on recommender systems has specifically begun addressing how loss aversion and nudging impact their optimal design (Paudel et al., 2018; Jesse and Jannach, 2021). Sun et al. (2022) design an algorithm that proactively incorporates predicted behavioral deviations in order to improve recommendations, while Kleinberg et al. (2022) study how the gap between System 1 and System 2 thinking impacts engagement optimization. Beyond recommender systems, Prospect Theory is explicitly considered in the collaboration of decision-makers with robots by Kwon et al. (2020) and with an AI e.g. by Ye et al. (2022).

The remaining article is structured as follows. Section 2 lays out a model of recommendation-dependent choices. Section 3 solves the model for a specific example and previews the general results in Section 4. In Section 5, we consider strategic non-recommendations within our model. Section 6 provides institutional and psychological foundations of our model. We discuss extensions in Section 7 before concluding in Section 8.

2 Setup

A decision-maker leverages information provided by a machine to take a decision $A \in \{\text{safe}, \text{risky}\}$ about an instance with outcomes $Y \in \{\text{good}, \text{bad}\}$. The decision-maker wants to take the safe decision when faced with a bad outcome, but prefers the risky decision when the outcome is good. For example, the decision-maker may be a judge who decides whether to release ($A = \text{risky}$) or jail ($A = \text{safe}$) a defendant, where the defendant may turn out to commit an offense or fail to appear ($Y = \text{bad}$) if released on bail or may appear without any new criminal activity ($Y = \text{good}$).

We assume that the decision-maker and machine both receive a signal X about the instance at hand that encodes any commonly known context (such as covariates) and information about the distribution of Y. In addition, the human decision-maker obtains a private signal H, which may include properties of the specific instance only visible in-person, and the machine receives a private signal M, which can encode information deduced from training data. For example, both the judge and their algorithmic assistant may have access to the sentencing history of a defendant, while the judge learns additional information from the defendant answering questions in court and the algorithm also synthesizes systematic insights from a large database of past defendants.

Jointly, the outcome Y and the signals X, H, M follow a (known) distribution P. Since the context X is known to both the decision-maker and the algorithm, we condition on it throughout, and consider the resulting joint distribution

$$(p_x(H,M) = \mathrm{P}(Y{=}\text{bad}|X{=}x, H, M), H, M) \mid X{=}x \sim \mathrm{P}_x \,. \tag{1}$$

In the judge example, P_x represents the distribution of the probability of not appearing or committing a new crime together with the private information the judge and the algorithm have about a defendant, holding jointly available information (such as the defendant's criminal record) fixed.

We assume that we aim to take an action A that minimize expected loss (risk) $\mathrm{E}[\ell(Y, A)]$ for the loss function

$$\ell(y,a) = \begin{cases} c_I, & y = \text{good}, a = \text{safe}, \\ c_{II}, & y = \text{bad}, a = \text{risky}, \end{cases} \tag{2}$$

where the two cases cover the two mistakes of choosing the safe option despite the outcome being good (leading to $c_I > 0$, Type-I error) or the risky decision in a bad case (leading to $c_{II} > 0$, Type-II error). For the jail decision, c_I is the cost of jailing a defendant who would not engage in criminal activity, and c_{II} the cost of releasing a defendant who commits a new crime or fails to re-appear.

In taking the action A, the decision-maker is helped by a simple recommendation $R = r_X(M)$ provided by the algorithm. We assume for tractability that this recommendation takes the form of a threshold rule[1]

$$r_x(m) = \begin{cases} \text{risky}, & \mathrm{P}(Y = \text{bad}|X{=}x, M{=}m) \leq \bar{q}_x, \\ \text{safe}, & \mathrm{P}(Y = \text{bad}|X{=}x, M{=}m) > \bar{q}_x. \end{cases} \tag{3}$$

Concretely, the algorithm in the jailing example would recommend the risky decision when its best prediction says that the bad outcome has a low probability, and would recommend the safe decision otherwise.

Binary recommendations represents a compression of information, and express the idea that the algorithmic recommendation is unable to preserve (or the decision-maker is unable to comprehend) the full complexity of the underlying information. In our model, we consider an extreme version of this compression, and assume that a binary recommendation is all the information that is conveyed by the algorithm. In Sections 5 and 7, we also consider the case where the algorithm provides more complex recommendations.

The human decision-maker takes an action $A = a_X(H, R) \in \{\text{safe}, \text{risky}\}$ that depends on their private information H and the recommendation R. The resulting (conditional) expected loss (risk) $\mathrm{E}_x[\ell(Y, A)]$ depends on the machine and human decisions through the realized decision, and incorporates the machine recommendation only indirectly through the information supplied to the decision-maker.

As a crucial deviation from standard analysis, we assume that the decision-maker experiences a decision loss $\ell^*(Y, A, R)$ that deviates from the consequence of the action alone. Specifically, we assume that they experience additional loss when they make a mistake that deviates from the machine recommendation,

$$\ell^*(y, a, r) = \ell(y, r) + \begin{cases} \Delta_I, & y = \text{good}, a = \text{safe}, r = \text{risky}, \\ \Delta_{II}, & y = \text{bad}, a = \text{risky}, r = \text{safe}, \end{cases} \tag{4}$$

with $\Delta_I, \Delta_{II} \geq 0$. Unlike the expectation of loss $\ell(Y, A)$, (conditional) expected decision loss $\mathrm{E}_x[\ell^*(Y, A, R)]$ now also depends on the recommendation beyond its information content. This ad-

[1] Threshold rules need not be optimal, but they represent a realistic reference point that simplify the analysis and provide a natural structure for reference dependence.

ditional loss can come from institutional features, such as when the judge is worried about reprisals for releasing a defendant who commits a crime against the recommendation of the algorithm. In Section 6, we also provide a micro-foundation in terms of a canonical model from behavioral science of reference-dependent preferences with loss aversion.

Throughout, we assume that the designer of the algorithm still aims to minimize expected (conditional) loss $\mathrm{E}_x[\ell(Y, A)]$ when choosing recommendations, while also anticipating that choices by the decision-maker now suffer from recommendation dependence. We therefore assume that the additional loss associated with deviations from the recommendation does not enter our final calculation of expected loss directly. In Section 7 we list the case where we care about (part of) this additional loss in designing the recommendations as a potential extension of our model.

3 Solution for a Simple Example

To drive intuition about the general solution to this problem, we first develop the solution to a simple instance of our setup. Consider the private signals, H and M, being drawn independently from a uniform distribution on $[0, 1]$. For the purposes of the example we will fix (and thus ignore) the context X. Let Y be deterministic in terms of H and M,

$$Y = \begin{cases} \text{bad}, & H + M \geq 1, \\ \text{good}, & \text{otherwise.} \end{cases}$$

In this section, we qualitatively analyze the properties of different decision rules and their relationship to underlying recommendations. All figures in this section show relative losses for $c_{II} > c_I$, so we assume that the cost of a risky action in the bad state is higher than that of a safe decision in the good state. When we introduce recommendation dependent utilities, we set $\Delta_I = 0$ (no *additional* cost of taking the safe action) while allowing $\Delta_{II} > 0$ (additional cost when deviating from the safe recommendation of the algorithm and making a mistake). In other words, we observe a situation where Type-II errors are worse than Type-I errors, and further, where the decision-maker perceives an additional penalty for a Type-II error if they deviate from the (safe) recommendation.

3.1 Oracle Action

Figure 1 provides a visual representation of Y as a function of to H and M. An oracle with access to both of these signals is able to achieve zero loss by always taking the corresponding optimal (and thus lossless) action

$$A^* = \begin{cases} \text{risky}, & Y = \text{good}, \\ \text{safe}, & Y = \text{bad.} \end{cases}$$

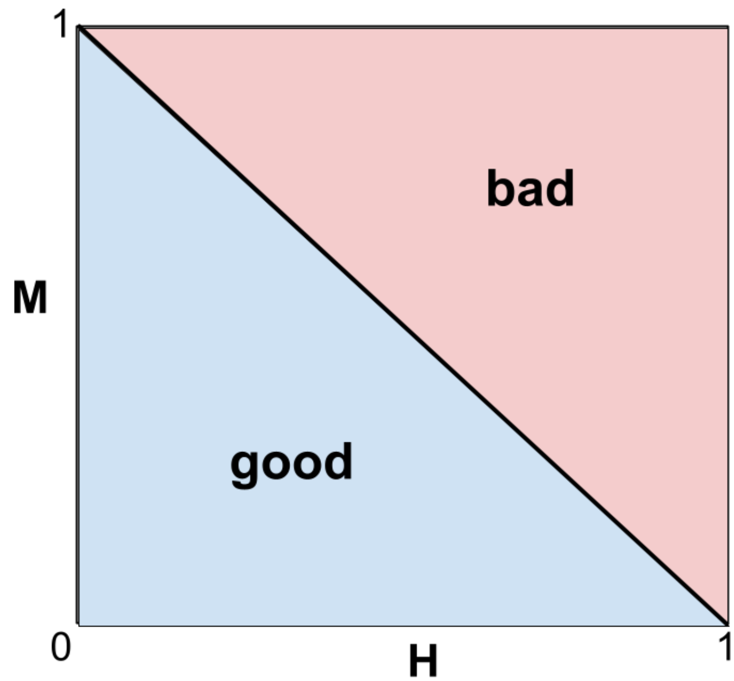


Figure 1: The signal space of $H \times M$ can be represented as the unit square. The oracle A^* takes the safe action in the upper right section where the label is bad, while it takes the risky action in the lower left section where the label is good.

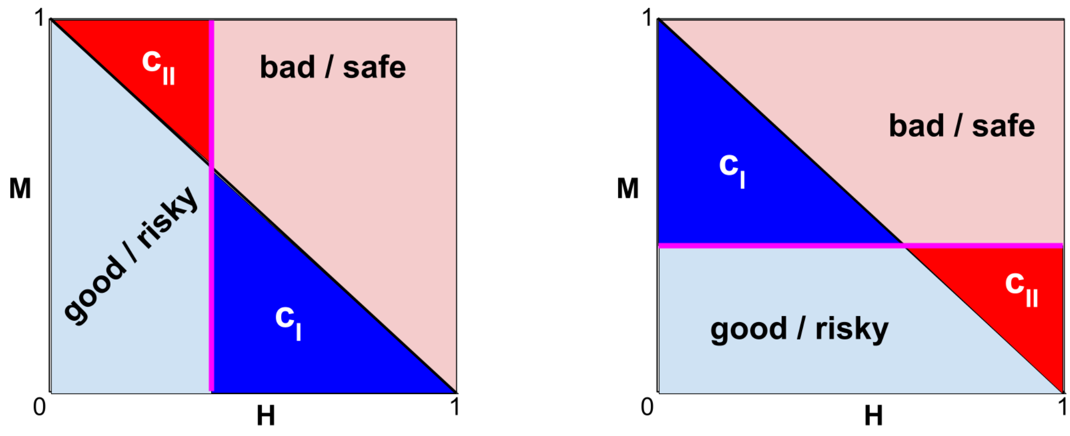
3.2 Individual Performance of Decision-Maker and Algorithm on Their Own

When the decision-maker and machine each take decisions by themselves, they need to take their decisions based solely on their observed private signal. Both $\mathrm{P}(Y{=}\text{bad}|H{=}h) = h$ and $\mathrm{P}(Y{=}\text{bad}|M{=}m) = m$ are monotonically increasing in h and m, respectively. Thus, both agents' optimal actions can be described in terms of threshold rules $A^h = a^h(H)$ and $A^M = a^m(M)$ on their signal with

$$a^h(h) = \begin{cases} \text{risky}, & h \leq \bar{p}^* = \frac{c_I}{c_I + c_{II}}, \\ \text{safe}, & h < \bar{p}^*, \end{cases} \qquad a^m(m) = \begin{cases} \text{risky}, & m \leq \bar{p}^*, \\ \text{safe}, & m > \bar{p}^*, \end{cases}$$

where the threshold $\bar{p}^* = \frac{c_I}{c_I + c_{II}}$ balances Type-I and Type-II errors optimally. We give a more detailed argument for the optimality of these actions in Appendix A.

Figure 2 shows the different signal combinations under which the human decision-maker and machine observe losses when utilizing threshold rules on their signals. In both cases, the threshold is set so that the marginal increase in Type-I costs is the same as the marginal decrease in Type-II



(a) Human acting alone

(b) Machine acting alone

Figure 2: Acting alone, both the human decision-maker and machine take their action according to a threshold rule based on their signal. Colors are used to represent the underlying outcomes (blue for good and red for bad). The dark regions represent losses, while the light regions are those where the chosen action agrees with the oracle.

costs when we shift the decision threshold up. Since each point in the figures is equally likely to occur, the losses are proportional to the area in the graph multiplied by the respective cost.

3.3 Recommendation without Recommendation Dependence

We next consider the case where the machine provides a recommendation rather than takes a decision. When the machine is moved upstream of the decision-maker in the decision process, the machine's output is utilized as a recommendation $R = r(M)$ rather than an action $A^m = a^m(M)$. Instead of attempting to minimize decision loss directly, the machine aims to choose a recommendation that leads to actions with low loss, and therefore optimally gives a recommendation that maximizes information relevant to the decision-maker.[2]

Figure 3 shows how the machine implementing a threshold recommendation affects the signal instances which lead to Type-I and Type-II errors once we consider the decision-maker response. For a machine threshold $\bar{q}$, the decision-maker now learns $M \in [0, \bar{q}]$ from $R = r(M) =$ risky and $M \in (\bar{q}, 1]$ from $R = r(M) =$ safe. The human decision-maker's optimal response to any threshold rule by the machine is to set two separate signal thresholds for its decision, one when $R = r(M) =$ safe and one when $R = r(M) =$ risky.

Given the recommendation R, in the baseline case without recommendation dependence the

[2]Were preferences misaligned between the human and machine (which we discuss in Section 7), the optimal recommendation would become more complicated as we would have to trade off information entropy and preference alignment (related to Kamenica and Gentzkow, 2011).

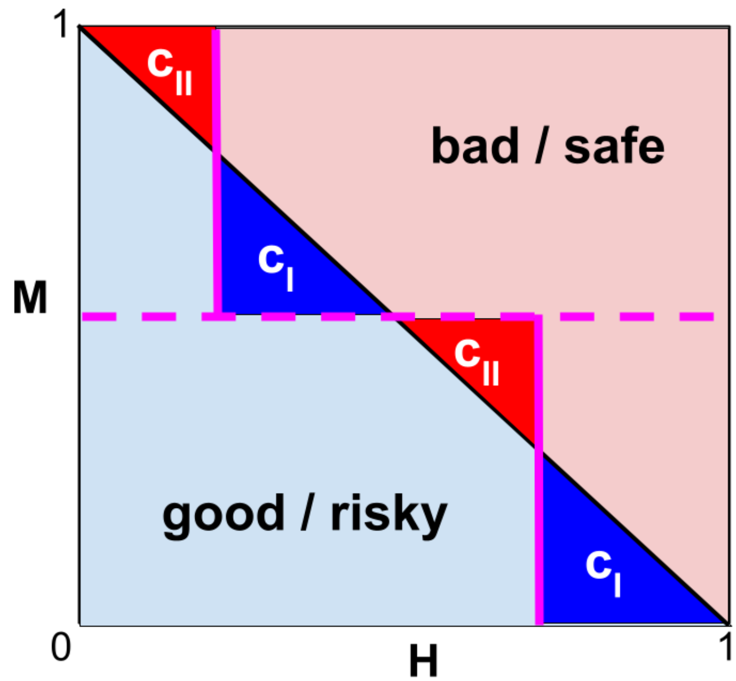


Figure 3: The machine's recommendation (here shown as a dotted pink line) creates two separate problems for the human decision-maker, one for each recommended action. In each the human sets a decision threshold to balance errors just as they did when working independently.

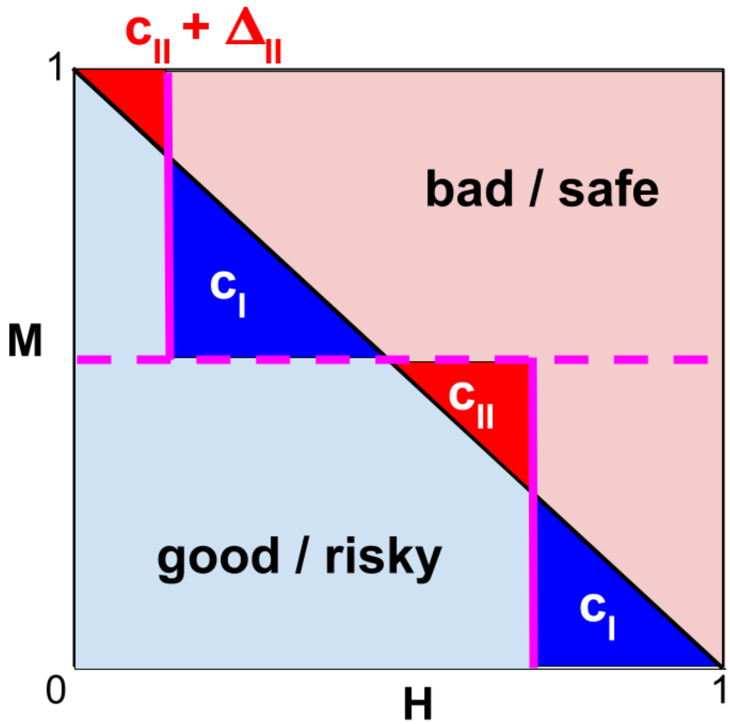


Figure 4: When the machine recommends $R =$ safe, the human takes $A =$ risky, and the outcome is $Y =$ bad, the human incurs additional decision loss. Thus, the humans signal threshold for taking $A =$ risky increases when $R =$ safe. This makes the baseline recommendation $R = R^\circ$ suboptimal because it does not efficiently trade off costs any more.

decision-maker optimally takes a decision $A = a(H, R)$ that can be written as a threshold rule in H with different thresholds for $R =$ risky and $R =$ safe, and replicates the divisions of $a^h(H)$ on $M \in [0, \bar{q}]$ and $M \in (\bar{q}, 1]$, respectively. By backwards induction, we can solve for the optimal recommendation in this baseline case, which maximizes informativeness by choosing a threshold $\bar{q}^\circ = \frac{1}{2}$ no matter the costs c_I, c_{II}. In optimum, we obtain recommendations $R^\circ = r^\circ(M)$ and actions $A^\circ = a^\circ(H, R^\circ)$ with

$$r^\circ(m) = \begin{cases} \text{safe}, & m \leq \frac{1}{2}, \\ \text{risky}, & m > \frac{1}{2}, \end{cases} \qquad a^\circ(h, r) = \begin{cases} \text{risky}, & h \leq \frac{\mathbb{1}(r=\text{risky})}{2} + \frac{c_I}{2(c_I + c_{II})}, \\ \text{safe}, & \text{otherwise}. \end{cases}$$

Specifically, the optimal recommendation R° differs from the optimal machine decision A^m except for the symmetrical edge case $c_I = c_{II}$.

3.4 Recommendation with Recommendation Dependence

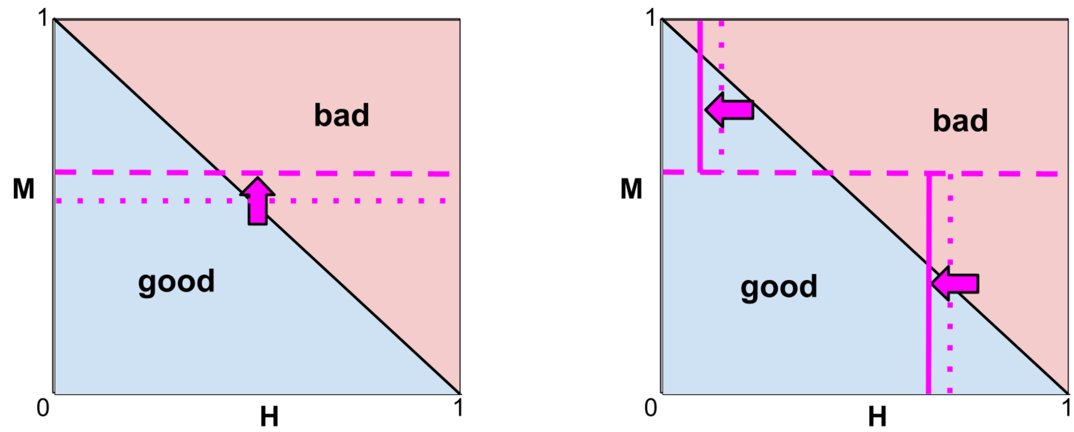
So far, we have considered a decision-maker who takes optimal decisions that minimize expected loss. We now consider a decision-maker who perceives additional reference-dependent decision loss $\Delta_{II} > 0$ whenever they take a risky decision $A =$ risky against a safe recommendation $R =$ safe when that decision turns out to be suboptimal, that is, when $Y =$ bad is realized.

Recommendation dependence creates a misalignment of decision loss between the human decision-maker and the machine whenever the recommendation $R =$ safe is given, leading to an over-adherence to that recommendation. As a result, the recommendation $R = R^{\circ}$ from the baseline case is not optimal any more since it creates an inefficient imbalance between Type-I and Type-II errors (Figure 4). The decision-maker is unaffected when $R =$ risky is recommended, but observes an increased (perceived) cost of a Type-II error to $c_{II} + \Delta_{II}$ when $R =$ safe. This causes the decision-maker to act overly cautious in the face of a safe recommendation.

In response to the inefficient choice by the decision-maker, the machine optimally reduces how often it offers a recommendation of $R =$ safe, opting to increase the probability of offering an $R =$ risky recommendation. While reducing the amount of information preserved in the recommendation, this shift increases the probability that the (sub-optimal) decision-maker takes the oracle decision.

In Figure 5, we illustrate how the optimal recommendation $R^{\triangle} = r^{\triangle}(M)$ with awareness of the resulting recommendation-dependent human decision $A^{\triangle} = a^{\triangle}(H, R^{\triangle})$ differs from the baseline case. In essence, the human's perceived loss, and thus their selections of actions, are misaligned with the loss-minimization objective of the machine when $R =$ safe is recommended. On the margin, the machine would then prefer the human to operate under the recommendation of $R =$ risky more often. The marginal return to shifting the threshold leads to an optimum in which Type-II errors are less frequent and Type-I errors contribute to more loss relative to the optimal division without recommendation dependence.

While we provide explicit expressions of optimal recommendations and resulting decisions in Appendix A, we summarize here some qualitative features of the solution. First, recommendation dependence increases adherence to the recommendation, which reduces the information revealed by the decision-maker when the distortionary recommendation (here, the safe recommendation) is given. Second, in response to this distortion, the optimal recommendation that takes recommendation dependence into account differs from the optimal recommendation without recommendation dependence, and suggests the distortionary recommendation less often. Finally, these effects get stronger the stronger the recommendation dependence is.



(a) The machine shifts its threshold in order to reduce the probability of the region with misaligned decision losses.

(b) In response the decision-maker becomes more likely to take the $A =$ risky decision in both recommendation regions.

Figure 5: Optimal decision thresholds are adjusted in response to recommendation dependence. The dotted pink lines show the thresholds in Figure 4, while the arrows depict the optimal change in machine threshold (left) and resulting adjustment of conditional decision-maker choices, which together balance Type-I and Type-II errors efficiently given reference-dependent choices (right).

4 General Solution

Having explored a simple example, we now consider the structure and consequences of human decisions given different algorithmic recommendations in the general model. We start by describing the optimal solutions an oracle would take, along with the decisions taken by the human decision-maker or the machine in isolation. We then describe the decision that is responsive to an algorithmic recommendation, and derive properties of the optimal recommendation threshold.

4.1 Reference Cases

As a baseline, we consider an oracle solution $A^* = a^*_X(H, M)$ of a decision-maker with full access to the H and M signals (and not exhibiting any reference dependence), where

$$a^*_x(h,m) = \begin{cases} \text{risky}, & p_x(h,m) \leq \bar{p}^* = \frac{c_I}{c_I + c_{II}}, \\ \text{safe}, & p_x(h,m) > \bar{p}^*, \end{cases}$$

yielding conditional expected loss $\mathrm{E}_x[\min((1 - p_x(H,M))\, c_I, p_x(H,M)\, c_{II})]$, which is minimal given the available information.[3] Two alternative benchmarks are the optimal decision $A^h = a^h_X(H)$ taken

[3] In principle, any choice at the threshold $p_x(h,m)$ is fine, and we make an arbitrary assumption about its direction here. As long as all distributions are continuous are the optimal rules still a.s. unique.

by the human decision-maker without machine input (and thus without reference dependence) and the optimal decision $A^m = a^m_X(M)$ by the machine directly without human interference, where

$$a^h_x(h) = \begin{cases} \text{risky}, & \mathrm{E}_x[p_x(h, M)|H{=}h] \leq \bar{p}^*, \\ \text{safe}, & \mathrm{E}_x[p_x(h, M)|H{=}h] > \bar{p}^*, \end{cases} \qquad a^m_x(m) = \begin{cases} \text{risky}, & \mathrm{E}_x[p_x(H, m)|M{=}m] \leq \bar{p}^*, \\ \text{safe}, & \mathrm{E}_x[p_x(H, m)|M{=}m] > \bar{p}^*. \end{cases}$$

In all of these cases, the optimal decisions are to choose the risky decision whenever the (posterior) probability of the bad outcome given all information available in the respective case is at most the critical, context-independent value $\bar{p}^*$. These choices minimize conditional expected loss $\mathrm{E}_x[\ell(Y, A)]$ given the respective information.

4.2 Recommendation-Dependent Decisions

We next consider the reactions of a decision-maker with reference-dependent preferences, represented by the offsets Δ_I, Δ_{II} (including the baseline case of $\Delta_I = 0 = \Delta_{II}$), to a recommendation $R = r_X(M)$ with threshold $\bar{q}_X$ as in (3). In this case, the decision-maker's optimal policy is $A^\triangle = a^\triangle_X(H, R)$ with

$$a^\triangle_x(h, r) = \begin{cases} \text{risky}, & \mathrm{E}_x[p_x(h, M)|H{=}h, r_x(M){=}r] \leq \bar{p}^r(\Delta_I, \Delta_{II}), \\ \text{safe}, & \mathrm{E}_x[p_x(h, M)|H{=}h, r_x(M){=}r] > \bar{p}^r(\Delta_I, \Delta_{II}), \end{cases} \tag{5}$$

$$\bar{p}^r(\Delta_I, \Delta_{II}) = \begin{cases} \frac{c_I + \Delta_I}{c_I + c_{II} + \Delta_I}, & r = \text{risky}, \\ \frac{c_I}{c_I + c_{II} + \Delta_{II}}, & r = \text{safe}. \end{cases} \tag{6}$$

This choice minimizes conditional expected decision loss $\mathrm{E}_x[\ell^*(Y, A, R)]$ given the recommendation policy $R = r_X(M)$, which is obtained from (3). The properties of the resulting decisions depend on the choice of threshold $\bar{q}_x$ for the recommendation. Here, we specifically consider the following three choices of recommendations:

1. The recommendation is chosen optimally for the case where it is implemented directly, $\bar{q}_x = \bar{p}^* = \frac{c_I}{c_I + c_{II}}$. In this case, $R = A^m$, as the machine recommends its own optimal decision.

2. The recommendation is chosen optimally for the case where the decision-maker chooses an action *rationally* in order to minimize expected loss, $\bar{q}^\circ_x \in \arg\min_{\bar{q}_x} \mathrm{E}_x \left[\min_a \mathrm{E}_x[\ell(Y, a)|r_x(M)]\right]$, leading to $R = R^\circ$. In this case, the optimal choice of recommendation threshold maximizes the information passed on to the decision-maker.

3. The recommendation is chosen optimally for the case where the decision-maker exhibits ref-

erence dependence. Now, the decision-maker chooses a threshold

$$\bar{q}_x^{\triangle}(\Delta_I, \Delta_{II}) \in \arg\min_{\bar{q}_x} \mathrm{E}_x \left[\ell(Y, a_x^{\triangle}(H, r_x(M))\right]$$

that minimizes expected loss, and (correctly) anticipates that the decision-maker instead minimizes expected recommendation-dependent decision loss $\mathrm{E}_x[\ell^*(Y, A, R)]$, leading to a recommendation $R = R^{\triangle}$. The optimal choice of threshold trades off the gain of information provision against the cost of loss distortion.

We note that $\bar{q}_x^{\circ} = \bar{q}_x^{\triangle}(0, 0)$, that is, the rational case is included as a special case of the recommendation-dependent case with $\Delta_I = 0 = \Delta_{II}$. Before we discuss the comparative statics of of optimal actions and recommendations in more detail, we note that the three thresholds are not generally the same.

Remark 1 (Optimal recommendation is not the same as optimal choice)**.** *In general, optimal recommendations are not the same as optimal decisions, even in the rational case ($\bar{q}_x^{\circ} \neq \bar{p}^*$ except in specific cases). In addition, recommendation dependence further changes the optimal threshold ($\bar{q}_x^{\triangle}(\Delta_I, \Delta_{II}) \neq \bar{q}_x^{\circ}$ except in specific cases or if $\Delta_I = 0 = \Delta_{II}$).*

We note that optimal recommendation thresholds can vary with the context x (as the structure and complementary of information changes), unlike optimal decision thresholds, which are always the same. Further, we assume here that the decision-maker correctly understands the recommendation. We consider cases in which the decision-maker makes simplified, possible wrong assumptions about the recommendation they are given in Section 7, including the case where the decision-maker naively (and falsely) assumes that they do not exhibit recommendation dependence and that the algorithm is optimized accordingly.

4.3 Consequences of Recommendation Dependence

We now discuss the qualitative and quantitative consequences of recommendation dependence on resulting decisions. First, recommendation dependence can mean that providing recommendations can make outcomes worse relative to the baseline case of not providing *any* recommendations, unless the recommendations are optimized for taking recommendation dependence into account.

Remark 2 (Comparison to choice without recommendation)**.** *With recommendation-dependent preferences, providing a recommendation can be worse than not providing a recommendation at all. At the same time, when $\Delta_I = 0$ (that is, there is no excess decision loss when the good outcome materializes), then an optimal choice $\bar{q}_x = \bar{q}_x^{\triangle}(\Delta_I, \Delta_{II})$ of the threshold improves expected loss relative to the case where no recommendation is provided.*

This stands in contrast to the case without recommendation-dependent preference, in which case any (correctly interpreted) recommendation (weakly) improves loss. The reason for the inefficiency is that increasing reference dependence means that the decision-maker follows the recommendation more than is optimal for minimizing expected loss. Specifically, as Δ_I and Δ_{II} increase, the probability that the action equals the recommendation increases, holding the recommendation policy fixed.

Proposition 1 (Recommendation dependence increases adherence)**.** *Holding the recommendation threshold $\bar{q}_x$ fixed, the probabilities* $\mathrm{P}_x(A^\triangle{=}R|R{=}\text{risky})$ *and* $\mathrm{P}_x(A^\triangle{=}R|R{=}\text{safe})$ *(weakly) increase with Δ_I and Δ_{II}, respectively.*

For a fixed recommendation, the level of adherence to the recommendation therefore becomes inefficient relative to the optimal level at $\Delta_I = 0 = \Delta_{II}$. The degree of this inefficiency generally depends on the strength of the signal available to the algorithm and the decision-maker for predicting the label of interest, with recommendation dependence having a larger effect for harder (more noisy) decisions. As one extreme, consider the case where the human decision-maker, after observing the recommendation, is sure about the label and can take the oracle action. In this case, there is no chance of an error, so the additional recommendation-dependent decision loss does not affect choices. On the other hand, if the probability of errors is large no matter the choice, then recommendation dependence may have an outsize effect on choices by making alignment with the algorithm the main driver of the decision.

4.4 Optimal Recommendations that Anticipate Recommendation Dependence

The above results have considered fixed recommendations. We next consider the case where recommendations are chosen optimally, and therefore change in response to increased recommendation dependence. In order to do so, we impose additional regularity conditions on the relevant distributions that will later allow us to express relevant comparative statics. Holding $X = x$ fixed and writing $Q = \mathrm{P}_x(Y = \text{bad}|M) = \mathrm{E}_x[p_x(H, M)|M]$ for the random variable that captures the best machine guess of the probability of the bad outcome, the machine recommendation for the threshold $\bar{q}_x$ are $R = \text{risky}$ for $Q \leq \bar{q}_x$ and $R = \text{safe}$ for $Q > \bar{q}_x$. Given the respective recommendation, the decision-maker observes the probabilities

$$P^{\text{risky}} = \mathrm{P}_x(Y = \text{bad}|H, Q \leq \bar{q}_x) = \mathrm{E}_x[p_x(H, M)|H, Q \leq \bar{q}_x],$$
$$P^{\text{safe}} = \mathrm{P}_x(Y = \text{bad}|H, Q > \bar{q}_x) = \mathrm{E}_x[p_x(H, M)|H, Q > \bar{q}_x],$$

respectively, and chooses $A = \text{risky}$ if and only if $P^R \leq \bar{p}^R$. For a given x, the distributions of $P^{\text{risky}}|Q \leq \bar{q}_x, X{=}x$ and $P^{\text{safe}}|Q > \bar{q}_x, X{=}x$ for different thresholds $\bar{q}$ as well as the distribution of $Q|X{=}x$ together describe the relevant properties of $(Y, H, M)|X{=}x$. As a first regularity assumption, we assume that these distributions are all described by a well-behaved family of densities.

Assumption 1 (Well-behaved conditional densities)**.** *Conditional on $X = x$, Q is absolutely continuous with respect to Lebesgue measure on* $[0, 1]$*, with a continuously differentiable density that is everywhere positive. Conditional on $X = x$ and all $\bar{q} \in [0, 1]$, $P^{\text{risky}}|Q \leq \bar{q}_x, X{=}x$ and $P^{\text{safe}}|Q > \bar{q}_x, X{=}x$ are absolutely continuous with respect to Lebesgue measure on* $[0, 1]$*, with continuous densities $f_x^{\text{risky}}(p; \bar{q}), f_x^{\text{safe}}(p; \bar{q})$, respectively. Furthermore, $f_x^{\text{risky}}(p; \bar{q}), f_x^{\text{safe}}(p; \bar{q})$ are twice continuously differentiable in $\bar{q}$.*

Under these regularity assumptions, we provide a first result on the relationship of recommendation dependence to optimal recommendations. Specifically, as recommendation dependence becomes dominant, the optimal decision thresholds reverts to that of the optimal decision taken by the machine directly, as adherence to the recommendation becomes perfect.

Proposition 2 (Optimal recommendation threshold reverts to optimal decision threshold)**.** *Under Assumption 1, as the severity of recommendation dependence increases ($\Delta_I \to \infty$ and $\Delta_{II} \to \infty$), the optimal recommendation threshold reverts to the optimal threshold of a decision taken directly by the machine $\bar{q}_x^{\triangle}(\Delta_I, \Delta_{II}) \to \bar{p}^*$.*

For this result, the conditions from Assumption 1 rule out certain degenerate cases where the decision-maker has certainty about the true label, and therefore does not respond to the cost of deviating from the recommendation. To obtain additional comparative statics for small changes in Δ_I, Δ_{II}, we make an additional monotonicity assumption.

Assumption 2 (Monotone net density)**.** *Conditional on $X = x$, $\mathrm{P}_x(Q{\leq}\bar{q}) \cdot f_x^{\text{risky}}(p; \bar{q})$ is monotonically increasing and $\mathrm{P}_x(Q{>}\bar{q}) \cdot f_x^{\text{safe}}(p; \bar{q})$ is monotonically decreasing in $\bar{q}$, for all $p \in [0, 1]$.*

This assumption expresses the idea that the conditional density of probabilities given the recommendation does not change by relatively more than the probability of the recommendation. Specifically, it requires that increasing the probability of receiving the recommendation $R = r$ increases the joint probability $\mathrm{P}_x(R = r, P^r \in [p, p + \mathrm{d}p])$ of receiving that recommendation and the probability being close to some value p, ruling out only cases in which the conditional distribution of probabilities changes drastically with a small change in the conditioning event. As an additional assumption, we assume that the optimal threshold is unique.

Assumption 3 (Unique optimal threshold)**.** *For all Δ_I, Δ_{II}, there is a unique optimal threshold $\bar{q}_x^{\triangle} = \bar{q}_x^{\triangle}(\Delta_I, \Delta_{II}) \in (0, 1)$, for which we also have that $\frac{\partial^2}{\partial \bar{q}_x^2} \mathrm{E}_x[\ell(Y, A^{\triangle})]|_{\bar{q}_x = \bar{q}_x^{\triangle}} > 0$.*

Here, the assumption that the second derivative is positive ensures that we obtain a locally unique loss-minimizing threshold and can describe its comparative statics via the relevant first-order condition. We can use these high-level assumptions to characterize general properties of recommendation-dependent actions in response to increased recommendation dependence. Our main results argues that an increase in reference dependence that applies to one of the recommendations decreases the probability of providing that recommendation.

Proposition 3 (Recommendation dependence shifts threshold)**.** *Under Assumptions 1, 2, and 3, the optimal threshold* $\bar{q}_x^{\triangle}(\Delta_I, \Delta_{II})$ *is (weakly) monotonically decreasing in* Δ_I *and (weakly) monotonically increasing in* Δ_{II}*.*

In particular, increasing the decision loss when the bad outcome materializes leads to a recommendation that is more likely to recommend the risky decision. The reason is that increased recommendation dependence in the case of a safe recommendation (higher Δ_{II}) means that the decision-maker does not take the risky decision enough. As an optimal response, the algorithm recommends the safe action less, thereby shifting away from the inefficient decision region.

5 Strategic Non-Recommendations

When recommendations distort choices, one solution is to strategically withhold recommendations in cases where the decision-maker knows better which decisions to take. Shashikumar et al. (2021) propose training a recommendation algorithm to return an "I don't know" response and apply the idea in the context of sepsis prediction. Within our formal model, we capture this idea by considering recommendations of the type

$$r_x(m) = \begin{cases} \text{risky}, & \mathrm{P}(Y = \text{bad}|X{=}x, M{=}m) \leq \bar{q}_x^{\text{low}}, \\ \text{don't know}, & \bar{q}_x^{\text{low}} < \mathrm{P}(Y = \text{bad}|X{=}x, M{=}m) \leq \bar{q}_x^{\text{high}}, \\ \text{safe}, & \mathrm{P}(Y = \text{bad}|X{=}x, M{=}m) > \bar{q}_x^{\text{high}}. \end{cases} \tag{7}$$

Such a recommendation structure relaxes the restriction that the provided information is binary to allow for three levels, so we would expect it to improve outcomes even in a model without recommendation dependence. However, with recommendation dependence, there can be an additional gain: if there is no additional cost from mistakes in the "don't know" case, then allowing for this third level also reduces the cost from recommendation dependence. In this sense, providing strategic non-recommendations has a higher benefit in our model relative to a rational baseline.

We first illustrate the gain from strategic non-recommendations in our example. Without considering recommendation dependence, adding a "don't know" option improves decisions by increasing the amount of information about the machine signal M preserved in the recommendation R. In the baseline case without recommendation dependence, the machine would optimally provide recommendations $R = r(M)$ based on thresholds $\bar{q}^{\text{low}} = \frac{1}{3}$ and $\bar{q}^{\text{high}} = \frac{2}{3}$, equally dividing the signal space in order to maximize the amount of information in the recommendation. Recommendation dependence changes the optimal signaling structure by reducing the frequency of situations in which $R = r(M)$ gives a safe signal, as this signal induces additional decision loss in the case of a bad outcome and thus distorts decisions. Thus, both thresholds will increase, as we detail in Appendix A. The resulting reduction in expected loss is larger than in the case without recom-

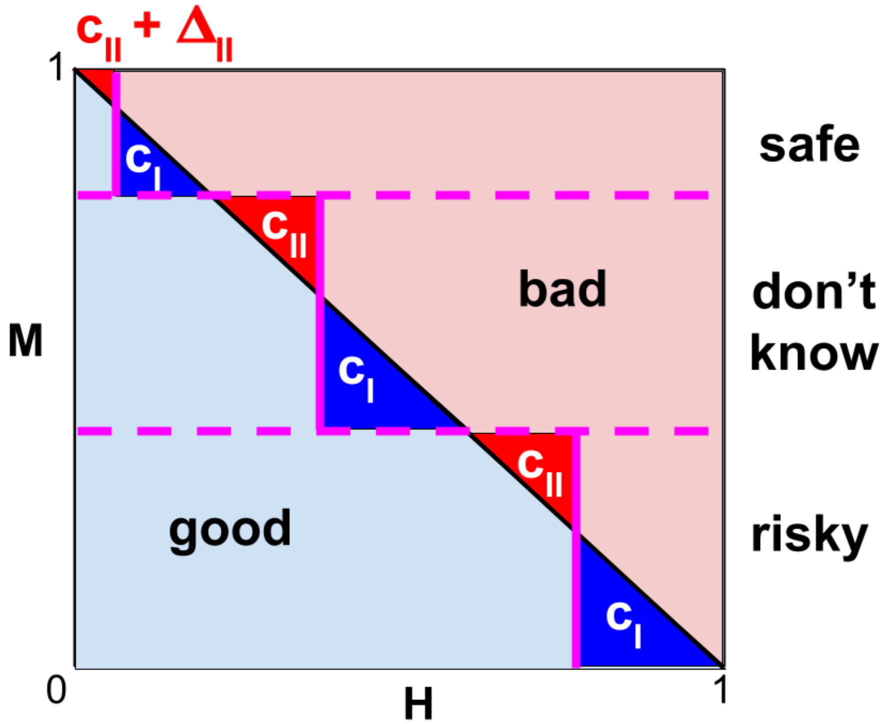


Figure 6: Strategic non-recommendation for the example developed in Section 3

mendation dependence.

Having discussed the effect of an additional recommendation option in the example, we now formalize the intuition that adding a "don't know" option is particularly relevant for the case of recommendation-dependent preferences in our general model from Section 4. In order to derive our general result, we denote by $\bar{q}_x^{\text{low}} = \bar{q}_x^{\text{low},\Delta}(\Delta_I, \Delta_{II}), \bar{q}_x^{\text{high}} = \bar{q}_x^{\text{high},\Delta}(\Delta_I, \Delta_{II})$ the optimal recommendation thresholds in (7) when the decision-maker takes recommendation-dependent choices with additional decision losses $\Delta_I, \Delta_{II} \geq 0$, and recall that $\bar{q}_x = \bar{q}_x^{\triangle}(\Delta_I, \Delta_{II})$ denotes the optimal threshold for the two-level recommendation case from Section 4. As a regularity assumption on $(p_X(H, M), H, M)$, we assume that these thresholds are ordered naturally.

Assumption 4 (Monotonic optimal thresholds)**.** *For all $\Delta_I, \Delta_{II} \geq 0$, the optimal thresholds fulfil $\bar{q}_x^{low} = \bar{q}_x^{low,\Delta}(\Delta_I, \Delta_{II}) \leq \bar{q}_x^{\triangle} = \bar{q}_x^{\triangle}(\Delta_I, \Delta_{II}) \leq \bar{q}_x^{high} = \bar{q}_x^{high,\Delta}(\Delta_I, \Delta_{II})$.*

Under the above assumptions, we now describe the improvement from adding the third "don't know" option. To this end, we compare the reduction in loss from the choice $A^{\triangle}$ based on an optimal recommendation with two options to the loss from the choice $A^{\blacktriangle}$ with an optimal three-level recommendation. By A° and $A^{\bullet}$ we denote the corresponding choices by a decision-maker who does not exhibit recommendation dependence, given recommendations with two and three levels, respectively, that are optimal in that case.

Proposition 4 (Gain from strategic non-recommendation)**.** *Under Assumptions 1, 2, and 4, the gain in loss from adding a third option is (weakly) larger in the case of reference-dependent choices, that is,* $\mathrm{E}_x[\ell(Y, A^{\triangle})] - \mathrm{E}_x[\ell(Y, A^{\blacktriangle})] \geq \mathrm{E}_x[\ell(Y, A^{\circ})] - \mathrm{E}_x[\ell(Y, A^{\bullet})]$ *for any* $\Delta_I, \Delta_{II} \geq 0$*.*

The proof shows that a stronger statement is true: the higher the degree of reference dependence, the larger the gain of adding a third option. Furthermore, the example provides a case where this inequality is strict.

6 Reasons and Evidence for Recommendation Dependence

In the previous two sections, we have explored the consequences of recommendation dependence on chosen actions and optimal design of the algorithm. Here, we discuss sources and evidence for our model of recommendation-dependent choices.

6.1 Institutional Factors

In many critical applications, negative outcomes can trigger additional scrutiny and formal audits. When recommendations are part of a decision process, ex-post suboptimal decisions that lead to undesirable outcomes may be seen as particularly problematic when they went against underlying recommendations. Doctors who are found to have caused medical harm with a procedure need to show their actions do not "deviate from accepted norms of practice in the medical community" to avoid a malpractice lawsuit (Bal, 2009). Deviations which lead to bad health outcomes are likely to draw additional scrutiny in this regard compared with physicians following an algorithmic standard. The specific outcomes of individual trials are not the basis for judge performance evaluations, but the reasoning of the judge's written opinions are (IAALS, 2022), which would likely address any deviation from recommended practice. Although evaluators are not supposed to consider the outcome when evaluating the judge's opinion, the evaluators have more information about the outcome than the judge did when they made their opinion. This may taint their perception of the logic used regardless of the conscious intent to do so. Hiring managers will be able to explain the hiring of an underperforming employee more easily if all indications of candidate quality are positive rather than if an algorithm or pre-employment evaluation had recommended against hiring the individual.

Institutional constraints may in many cases also predict an asymmetry in the penalty associated with taking ex-post suboptimal decisions that deviate from recommendations. Deviations which lead to safe decisions against risky advice may not be seen as equally problematic as risky decisions against safe advice, even if they are ex-post suboptimal. Doctors who order an extra test which was unnecessary may face some penalty from an insurer, but are unlikely to face a large outcry or malpractice lawsuit. A judge who jails a defendant who turns out to be of low risk of committing a new crime or failing to appear may not face scrutiny because the behavior outside jail is never observed. When a good applicant is not hired, there may be limited repercussions for the manager since performance is not visible.

6.2 Reference-Dependent Preferences with Loss Aversion

So far, we have mainly considered institutional justifications for considering recommendation dependence. In this section, we instead consider a psychological motivation from behavioral science that similarly yield that recommendations do not only affect decisions through the provision of information, but also by affecting decision utility. Specifically, we derive the specific decision loss ℓ^* in (4) from Prospect Theory (Kahneman and Tversky, 1979; Tversky and Kahneman, 1992), which has been one of the most established frameworks for describing systematic deviations from rational utility theory in behavioral economics and its applications (see e.g. Barberis, 2013, for an overview and assessment). Like Kleinberg et al. (2022), we therefore assume that there is a gap between welfare-relevant utility and the decision-maker's perceived utility when taking the decision.

We consider two central tenets of Prospect Theory to the decision-makers choice between safe and risky actions. First, we assume that choices are evaluated relative to a reference point, which we here assume is induced by the action R recommended by the algorithm. This means that the decision-maker evaluates losses relative to the reference loss $\ell(Y, R)$ that they would achieve if they followed the recommendation.

The second aspect of Prospect Theory we adopt to our setting is that the decision-maker puts more emphasis on losses relative to the reference point than on gains. Specifically, we assume that loss aversion takes the form of a factor $\lambda > 1$ by which losses are multiplied. This means that decision loss from outcome y relative to the reference point $\ell(y, r)$ from taking action a given recommendation r is given by

$$\ell^{\mathrm{PT}}(y, r, a) = \lambda[\ell(y, a) - \ell(y, r)]_+ - \ [\ell(y, a) - \ell(y, r)]_-,$$

where by $[\cdot]_+$ and $[\cdot]_-$ we denote the (absolute value of the) positive and negative parts, respectively. For the specific loss function from (2), the Prospect-Theory loss takes the form of a recommendation-dependent decision loss from (4).

Proposition 5 (Derivation from Prospect Theory)**.** *Decision-maker choices according to ℓ^{PT} are equivalent to choices according to ℓ^* with $\Delta_I = (\lambda - 1)c_I, \Delta_{II} = (\lambda - 1)c_{II}$.*

We note that this justification implies additional structure relative to the ad-hoc construction of losses in Section 2. Specifically, additional costs are larger in the case where the baseline cost of an error is larger. In the canonical case where taking the risky decision in the bad case has higher cost ($c_{II} > c_I$), this model of behavioral decision-making justifies a focus on the case with large Δ_{II}, where the decision-maker tends to be too cautious in response to a safe recommendation, and optimal recommendation thresholds should therefore lead to risky recommendations more often than in the reference case of rational choice.

In this section, we have been assuming that the machine recommendation represents a reference point relative to which the decision-maker evaluates their choices. While the recommendation itself represents a natural starting point, we could instead assume that the reference point is itself a result of beliefs about losses. Koszegi and Rabin (2006) propose a model in which the reference point (in loss space) is formed by the expected loss of the decision, where reference point and reference-dependent choices are then determined in equilibrium. We believe that closing the model in this way represents an attractive alternative approach to incorporating reference dependence through Prospect Theory into our approach.

6.3 Empirical Evidence

Here, we briefly mention emerging empirical evidence that suggests that human decisions depend on machine recommendations not just through the information provided. The empirical literature on algorithm-assisted human decision-making (e.g. Lai et al., 2021) has made observations that are particularly hard to align with a rational model of decision-making and become more reasonable when recommendation dependence is considered. We note, however, that in many of these cases it is challenging to separate evidence for recommendation dependence from evidence for automation bias, overconfidence in algorithms, anchoring, and the salience of information, since one of the most prominent consequence of these behavioral models is that the recommended actions is taken relatively more often.

In a model of rational and aligned choice, decision-makers always consider costless information, yet Snyder et al. (2022) observe participants preferring not to accept costless advice in an experiment. Under recommendation dependence, this behavior may stem from (sophisticated) users to perceiving higher loss and preferring top forgo recommendations. Fügener et al. (2021) hypothesize and demonstrate empirically that excess coordination due to algorithmic advice may destroy unique knowledge and reduce performance in an aggregated 'wisdom of the crowds' scenario. Stevenson and Doleac (2019) analyze the introduction of an algorithmic risk assessment tool that assists judges in sentencing decisions. They show that the introduction of the tool moves judge decisions in the direction suggested by the algorithm, but estimate that the changed behavior does not significantly affect the overall quality of the decisions in terms of incarceration rates, sentence lengths, and recidivism rates. Banker and Khetani (2019) document cases of over-dependence on algorithmic recommendations across multiple experiments in which algorithmic advice pushes human decision-makers towards making inferior, dominated choices. Green and Chen (2019b) and Fogliato et al. (2022a) provide evidence for anchoring effects, where adherence to algorithmic recommendations is larger when these are revealed initially rather than after eliciting provisional human judgements.

7 Extensions

To close this article, we briefly mention relevant extensions to the baseline models and findings.

7.1 Alignment of Preferences

We have assumed throughout that decision-maker and algorithm agree on their costs c_I and c_{II} of making mistakes, and only differ with respect to recommendation-dependent losses of the decision-maker. If the baseline costs c_I, c_{II} are already misaligned, we face a persuasion problem similar to Kamenica and Gentzkow (2011). In this case, recommendation dependence may improve decisions by increasing adherence to the preferred action of the algorithm designer, even if it comes at the cost of reducing revealed information.

We have also assumed that the additional loss associated with deviating from recommendations only affects the perceived loss of the decision-maker, and not directly the loss of the designer of the algorithm. As an alternative extension to our model, we could also assume that the designer aims to minimize (part of) this additional loss. This modification would change optimal thresholds. In the case where the designer of the algorithm fully incorporates the decision-makers perceived loss, choices are now perfectly aligned, but the additional cost associated with deviations from recommendations means that the loss of the designer is directly affected by costs from recommendations that the decision-maker does not follow through on.

7.2 Simple Cost from Deviation

In our main model, we assume that there are additional costs Δ_I, Δ_{II} affecting the decision loss that only come from (expected) Type-I and Type-II errors (when deviating from the recommendation). Here, we instead consider the case where any deviation from the recommendation is perceived as costly by the decision-maker, no matter whether it leads to errors or not. Assuming that there is a cost (in addition to expected loss $\mathrm{E}_x[\ell(Y, A)]$) of d^{risky} of deviating from the risky recommendation $R =$ risky and d^{safe} of deviating from the safe recommendation $R =$ safe, the resulting optimal decision is the same as (5) with thresholds

$$\bar{p}^r = \begin{cases} \min\left(\frac{c_I + d^{\text{risky}}}{c_I + c_{II}}, 1\right), & r = \text{risky}, \\ \max\left(\frac{c_I - d^{\text{safe}}}{c_I + c_{II}}, 0\right), & r = \text{safe}. \end{cases}$$

Our main results and comparative statics therefore still apply since the costs shift the recommendation-specific thresholds similarly to $\bar{p}^r$ from (6). However, relative to our baseline model, there are now cases where the decision-maker may go with the recommendation even when they know with certainty that it leads to an error. While this prediction may appear less realistic for modelling application areas and behavioral effects like those discussed in Section 6, it captures cases where a designer imposes a deviation cost on the decision-maker irrespective of actual outcome.

7.3 More Complex Recommendations

We have assumed that the provided information consists of binary recommendations only. While we could expand the model to more complex recommendations, we could instead consider a model where a probability forecast is provided *along* with a recommendation. In the case where risk preferences are misaligned, such a method could be justified by partly aligning preferences.

7.4 Mis-Interpretation by the Decision-Maker

Throughout, we have assumed that the decision-maker is able to interpret recommendations correctly. But in practice, the decision-maker may have a hypothesis about the recommendation that may not be fully accurate. As an extension, three approaches may be particularly relevant. The first is that the decision-maker assumes that the recommendation is an optimal machine decision, $R = A^m$. The second approach considers a decision-maker who is naive about their own reference dependence, so they assume that the recommendation is optimal for the case where they do not exhibit recommendation dependence, $R = R^\circ$. The third approach would be one where we assume that the decision-maker can only understand a simple representation or explanation of the recommendation.

7.5 Context-Sensitive Algorithmic Delegation to the Human Decision-Maker

We have so far considered a decision-maker who retains full decision authority and takes a decision based on a recommendation. Raghu et al. (2019) instead consider an algorithm that decides whether to delegate a decision to a human decision-maker, an optimization task they dub the "algorithmic automation problem." Within our framework, we could adapt the setting in Section 5, where instead of returning a "don't know" level the algorithm delegates decisions when it is unsure about the optimal action. Specifically, we could consider a hierarchical model of joint decision-making where the algorithm decides to choose an action directly or to delegate to the human decision-maker by taking a $D = d_X(M) \in \{\text{risky}, \text{delegate}, \text{safe}\}$ decision, such as

$$d_x(m) = \begin{cases} \text{risky}, & \mathrm{P}(Y = \text{bad}|X{=}x, M{=}m) \leq \bar{q}_x^{\text{low}}, \\ \text{delegate}, & \bar{q}_x^{\text{low}} < \mathrm{P}(Y = \text{bad}|X{=}x, M{=}m) \leq \bar{q}_x^{\text{high}}, \\ \text{safe}, & \mathrm{P}(Y = \text{bad}|X{=}x, M{=}m) > \bar{q}_x^{\text{high}}. \end{cases}$$

The action $A = a_x(M, D)$ would then be chosen by the decision-maker only for $D = \text{delegate}$, and otherwise equal to the respective machine decision, $A = D$ for $D \neq \text{delegate}$. In the context of our model, such a hierarchical decision would be obtained from the recommendation in (7) if Δ_I, Δ_{II}

are very high and the human decision-maker is never sure that the recommendation is the wrong decision (e.g. under the conditions of Proposition 2), since the decision-maker would never deviate from the recommended action in this case.

7.6 Optimal Recommendations Beyond Threshold Rules

For simplicity, we have so far considered recommendations that can be represented as simple threshold rules with respect to the probability forecast of the machine, specifically those in (3) and (7). These threshold rules are not necessarily optimal even among two- and three-level recommendations given by the machine. While we believe that such threshold rules are realistic in many cases and may be better understood by a human decision-maker than more complex rules, relevant extensions to our work include the derivation of generally optimal recommendations and of tractable sufficient conditions under which simple threshold rules are optimal.

7.7 Low-Level Sufficient Conditions for Comparative Statics

In order to derive comparative statics with respect to the effect of increased deviation costs in Proposition 3 and Proposition 4, we currently rely on the high-level Assumptions 2, 3, and 4. While we believe that they apply in many cases, we think that developing more applicable and economically meaningful sufficient conditions is an important next step in this project.

8 Conclusion

When we provide a decision-maker with a recommendations, they may not only react to its information content, but also see it as a default action that affects their preferences. In this article, we illustrate in a simple example and with general results how recommendation-dependent preferences create inefficiencies and affect the design of optimal recommendations. Our model suggests practically implementable modifications that reduce distortions by strategically altering or even withholding recommendations for instances where they may otherwise hurt more than they help. With our work, we hope to provide an example for the integration of more realistic models of human behavior into the design of algorithms, and hope that it can contribute to improving human–AI interaction in critical applications.

References

Athey, S. C., Bryan, K. A., and Gans, J. S. (2020). The Allocation of Decision Authority to Human and Artificial Intelligence. AEA Papers and Proceedings, 110:80–84.

Bai, B., Dai, H., Zhang, D., Zhang, F., and Hu, H. (2021). The Impacts of Algorithmic Work Assignment on Fairness Perceptions and Productivity. Academy of Management Proceedings, 2021(1):12335.

Bal, B. S. (2009). An Introduction to Medical Malpractice in the United States. Clinical Orthopaedics and Related Research, 467(2):339–347.

Banker, S. and Khetani, S. (2019). Algorithm overdependence: How the use of algorithmic recommendation systems can increase risks to consumer well-being. Journal of Public Policy & Marketing, 38(4):500–515.

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., and Horvitz, E. (2019a). Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 7:2–11.

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., and Horvitz, E. (2019b). Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):2429–2437.

Barberis, N. C. (2013). Thirty years of prospect theory in economics: A review and assessment. Journal of Economic Perspectives, 27(1):173–96.

Bastani, H., Bastani, O., and Sinchaisri, W. P. (2021). Improving human decision-making with machine learning. arXiv:2108.08454.

Buçinca, Z., Malaya, M. B., and Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1):188:1–188:21.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1):114.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2018). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. Management Science, 64(3):1155–1170.

Donahue, K., Chouldechova, A., and Kenthapadi, K. (2022). Human-Algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness. arXiv:2202.08821.

Donohue, K., Özer, O., and Zheng, Y. (2020). Behavioral Operations: Past, Present, and Future. Manufacturing & Service Operations Management, 22(1):191–202.

Fogliato, R., Chappidi, S., Lungren, M., Fisher, P., Wilson, D., Fitzke, M., Parkinson, M., Horvitz, E., Inkpen, K., and Nushi, B. (2022a). Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22, pages 1362–1374.

Fogliato, R., De-Arteaga, M., and Chouldechova, A. (2022b). A Case for Humans-in-the-Loop: Decisions in the Presence of Misestimated Algorithmic Scores. SSRN 4050125.

Fügener, A., Grahl, J., Gupta, A., and Ketter, W. (2021). Will Humans-in-The-Loop Become Borgs? Merits and Pitfalls of Working with AI. SSRN 3879937.

Gillis, T., McLaughlin, B., and Spiess, J. (2021). On the Fairness of Machine-Assisted Human Decisions. arXiv:2110.15310.

Green, B. and Chen, Y. (2019a). Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pages 90–99.

Green, B. and Chen, Y. (2019b). The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW):1–24.

Hampshire, R. C., Bao, S., Lasecki, W. S., Daw, A., and Pender, J. (2020). Beyond safety drivers: Applying air traffic control principles to support the deployment of driverless vehicles. PLOS ONE, 15(5):e0232837.

IAALS (2022). Judicial Performance Evaluation 2.0. Institute for the Advancement of the American Legal System.

Ibrahim, R., Kim, S.-H., and Tong, J. (2021). Eliciting Human Judgment for Prediction Algorithms. Management Science, 67(4):2314–2325.

Imai, K., Jiang, Z., Greiner, J., Halen, R., and Shin, S. (2020). Experimental Evaluation of Algorithm-Assisted Human Decision-Making: Application to Pretrial Public Safety Assessment. arXiv:2012.02845.

Jesse, M. and Jannach, D. (2021). Digital nudging with recommender systems: Survey and future directions. Computers in Human Behavior Reports, 3:100052.

Johnsen, L. C., Voigt, G., and Corbett, C. J. (2019). Behavioral Contract Design Under Asymmetric Forecast Information. Decision Sciences, 50(4):786–815.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47(2):263–292.

Kamenica, E. and Gentzkow, M. (2011). Bayesian Persuasion. American Economic Review, 101(6):2590–2615.

Kiani, A., Uyumazturk, B., Rajpurkar, P., Wang, A., Gao, R., Jones, E., Yu, Y., Langlotz, C. P., Ball, R. L., Montine, T. J., Martin, B. A., Berry, G. J., Ozawa, M. G., Hazard, F. K., Brown, R. A., Chen, S. B., Wood, M., Allard, L. S., Ylagan, L., Ng, A. Y., and Shen, J. (2020). Impact of a deep learning assistant on the histopathologic classification of liver cancer. NPJ Digital Medicine, 3(1):1–8.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2022). The Challenge of Understanding What Users Want: Inconsistent Preferences and Engagement Optimization. arXiv:2202.11776.

Koszegi, B. and Rabin, M. (2006). A Model of Reference-Dependent Preferences. The Quarterly Journal of Economics, 121(4):1133–1165.

Kwon, M., Biyik, E., Talati, A., Bhasin, K., Losey, D. P., and Sadigh, D. (2020). When humans aren't optimal: Robots that collaborate with risk-aware humans. In 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pages 43–52. IEEE.

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., and Tan, C. (2021). Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. arXiv:2112.11471.

Lakkaraju, H. and Bastani, O. (2020). "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. In Markham, A. N., Powles, J., Walsh, T., and Washington, A. L., editors, AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020, pages 79–85.

Lawrence, M., Goodwin, P., O'Connor, M., and Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. International Journal of Forecasting, 22(3):493–518.

Lima, G., Grgić-Hlača, N., Jeong, J. K., and Cha, M. (2022). The Conflict Between Explainable and Accountable Decision-Making Algorithms. 2022 ACM Conference on Fairness, Accountability, and Transparency, pages 2103–2113.

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151:90–103.

McGrath, S., Mehta, P., Zytek, A., Lage, I., and Lakkaraju, H. (2020). When Does Uncertainty Matter? Understanding the Impact of Predictive Uncertainty in ML Assisted Decision Making. arXiv:2011.06167.

Murray, N. M., Unberath, M., Hager, G. D., and Hui, F. K. (2020). Artificial intelligence to diagnose ischemic stroke and identify large vessel occlusions: a systematic review. Journal of NeuroInterventional Surgery, 12(2):156–164.

Palley, A. B. and Soll, J. B. (2019). Extracting the wisdom of crowds when information is shared. Management Science, 65(5):2291–2309.

Paudel, B., Luck, S., and Bernstein, A. (2018). Loss Aversion in Recommender Systems: Utilizing Negative User Preference to Improve Recommendation Quality. arXiv:1812.11422.

Raghu, M., Blumer, K., Corrado, G., Kleinberg, J., Obermeyer, Z., and Mullainathan, S. (2019). The Algorithmic Automation Problem: Prediction, Triage, and Human Effort. arXiv:1903.12220.

Schweitzer, M. E. and Cachon, G. P. (2000). Decision Bias in the Newsvendor Problem with a Known Demand Distribution: Experimental Evidence. Management Science, 46(3):404–420.

Shashikumar, S. P., Wardi, G., Malhotra, A., and Nemati, S. (2021). Artificial intelligence sepsis prediction algorithm learns to say "I don't know". NPJ Digital Medicine, 4(1):1–9.

Snyder, C., Keppler, S., and Leider, S. (2022). Algorithm Reliance Under Pressure: The Effect of Customer Load on Service Workers. SSRN 4066823.

Stevenson, M. T. and Doleac, J. L. (2019). Algorithmic Risk Assessment in the Hands of Humans. SSRN 3489440.

Su, X. (2008). Bounded Rationality in Newsvendor Models. Manufacturing & Service Operations Management, 10(4):566–589.

Sun, J., Zhang, D. J., Hu, H., and Van Mieghem, J. A. (2022). Predicting Human Discretion to Adjust Algorithmic Prescription: A Large-Scale Field Experiment in Warehouse Operations. Management Science, 68(2):846–865.

Taudien, A., Fügener, A., Gupta, A., and Ketter, W. (2022). The Effect of AI Advice on Human Confidence in Decision-Making. In Proceedings of the 55th Hawaii International Conference on System Sciences.

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. Journal of Risk and uncertainty, 5(4):297–323.

Ye, W., Bullo, F., Friedkin, N., and Singh, A. K. (2022). Modeling human-ai team decision making. arXiv:2201.02759.

A Solution of the Simple Example

Here we expand upon the solution to the simple example presented in Section 3. For these examples, it will be helpful to note that from Figure 1 we directly obtain that

$$\begin{aligned} \mathrm{P}(Y{=}\mathrm{bad}|H{=}h) &= h, \\ \mathrm{P}(Y{=}\mathrm{bad}|M{=}m) &= m, \end{aligned} \qquad \mathrm{P}(Y{=}\mathrm{bad}|H{=}h, M{\in}(\underline{m},\overline{m}]) = \begin{cases} 1, & h \geq 1-\underline{m}, \\ \frac{h-1+\overline{m}}{\overline{m}-\underline{m}}, & h \in [1{-}\overline{m}, 1{-}\underline{m}), \\ 0, & h < 1-\overline{m}. \end{cases} \tag{8}$$

We also note that in general, an optimal $\mathcal{F}$-measurable decision A that minimizes expected loss $\mathrm{E}[c_I \mathbb{1}(A = \text{safe}, Y = \text{good}) + c_{II}\mathbb{1}(A = \text{risky}, Y = \text{bad})]$ is given by

$$A = \begin{cases} \text{risky}, & \mathrm{P}(Y = \text{bad}|\mathcal{F}) \leq \bar{p}^* = \frac{c_I}{c_I + c_{II}}, \\ \text{safe}, & \mathrm{P}(Y = \text{bad}|\mathcal{F}) > \bar{p}^*, \end{cases} \tag{9}$$

and it is unique on $\mathrm{P}(Y = \text{bad}|\mathcal{F}) \neq \bar{p}^*$.

A.1 Individually Optimal Decisions

From (8) and (9) it directly follows that

$$a^h(h) = \begin{cases} \text{risky}, & h \leq \frac{c_I}{c_I + c_{II}}, \\ \text{safe}, & \text{otherwise}, \end{cases} \qquad a^m(m) = \begin{cases} \text{risky}, & m \leq \frac{c_I}{c_I + c_{II}}, \\ \text{safe}, & \text{otherwise}, \end{cases}$$

which are the a.s. unique optimal decisions for the respective optimal decisions.

A.2 Recommendation-Dependent Response to Threshold Recommendation

In response to a threshold recommendation $R = r^{\bar{q}}(M)$ where

$$r^{\bar{q}}(m) = \begin{cases} \text{risky}, & m \leq \bar{q}, \\ \text{safe}, & m > \bar{q} \end{cases}$$

with threshold $\bar{q}$, the human chooses an optimal decision as in (9) separately on $M \in [0, 1 - \bar{q}]$ with c_{II} replaced by $c_{II} + \Delta_{II}$ if given a recommendation of $R = \text{safe}$ and on $M \in [1 - \bar{q}, 1]$ if given the recommendation $R = \text{risky}$. A visual representation is shown in Figure 4. Since

$$\mathrm{P}(Y{=}\mathrm{bad}|H{=}h, M{\leq}\bar{q}) = \max\left(\frac{h-1+\bar{q}}{\bar{q}}, 0\right), \qquad \mathrm{P}(Y{=}\mathrm{bad}|H{=}h, M{>}\bar{q}) = \min\left(\frac{h}{1-\bar{q}}, 1\right)$$

from (8), with (9) it follows that the optimal response is $A = a^{\bar{q}}(H, R)$ with

$$a^{\bar{q}}(h,r) = \begin{cases} \text{risky}, & h \le \bar{h}^{\bar{q}}(r), \\ \text{safe}, & h > \bar{h}^{\bar{q}}(r), \end{cases}$$

with different thresholds $\bar{h}^{\bar{q}}(\text{risky}), \bar{h}^{\bar{q}}(\text{safe})$ that depend on the recommendation by

$$\begin{aligned}
\bar{h}^{\bar{q}}(r) &= (1-\bar{q})\mathbb{1}(r=\text{risky}) + ((1-\bar{q})\mathbb{1}(r=\text{safe}) + \bar{q}\mathbb{1}(r=\text{risky}))\,\frac{c_I}{c_I + c_{II} + \Delta_{II}\mathbb{1}(r=\text{safe})} \\
&= \frac{c_I}{c_I + c_{II} + \Delta_{II}\mathbb{1}(r=\text{safe})} + \begin{cases} (1-\bar{q})\frac{c_{II}}{c_I+c_{II}}, & r = \text{risky}, \\ -\bar{q}\frac{c_I}{c_I+c_{II}+\Delta_{II}}, & r = \text{safe}, \end{cases} \\
&= \bar{p}^* + \begin{cases} (1-\bar{q})\frac{c_{II}}{c_I+c_{II}}, & r = \text{risky}, \\ -\left(\bar{q} + \frac{\Delta_{II}}{c_I+c_{II}}\right)\frac{c_I}{c_I+c_{II}+\Delta_{II}}, & r = \text{safe}. \end{cases}
\end{aligned}$$

In particular, the decision-makers threshold in response to a recommendation of $R =$ risky will always be higher than that in response to $R =$ safe.

A.3 Optimal Threshold with Recommendation Dependence

Now that we have described the human action in response to a recommendation with threshold $\bar{q}$, we solve for the threshold $\bar{q}^{\triangle}$ that minimizes the expected loss

$$\begin{aligned}
&\mathrm{E}[\ell(Y, a^{\bar{q}}(H, r^{\bar{q}}(M)))] \\
&= c_I(\mathrm{P}(Y{=}\text{good}, a^{\bar{q}}(H,\text{safe})){=}\text{safe}, r^{\bar{q}}(M){=}\text{safe}) + \mathrm{P}(Y{=}\text{good}, a^{\bar{q}}(H,\text{risky})){=}\text{safe}, r^{\bar{q}}(M){=}\text{risky})) \\
&\quad + c_{II}(\mathrm{P}(Y{=}\text{bad}, a^{\bar{q}}(H,\text{safe})){=}\text{risky}, r^{\bar{q}}(M){=}\text{safe}) + \mathrm{P}(Y{=}\text{bad}, a^{\bar{q}}(H,\text{risky})){=}\text{risky}, r^{\bar{q}}(M){=}\text{risky})) \\
&= c_I \frac{\left((1{-}\bar{q})\,\frac{c_{II}+\Delta_{II}}{c_I+c_{II}+\Delta_{II}}\right)^2 + \left(\bar{q}\left(\frac{c_{II}}{c_I+c_{II}}\right)\right)^2}{2} + c_{II}\frac{\left((1{-}\bar{q})\frac{c_I}{c_I+c_{II}+\Delta_{II}}\right)^2 + \left(\bar{q}\left(\frac{c_I}{c_I+c_{II}}\right)\right)^2}{2} \\
&= \frac{c_I c_{II}}{c_I + c_{II}}\bar{q}^2 + \frac{c_I\left((c_{II}+\Delta_{II})^2 + c_I c_{II}\right)}{(c_I+c_{II}+\Delta_{II})^2}(1-\bar{q})^2,
\end{aligned}$$

corresponding to the respective triangles in Figure 4. Here, we can calculate the probabilities over the joint independent uniform as the areas of the triangles. We obtain the optimal threshold $\bar{q}^{\triangle}$ and optimal recommendation $R = r^{\triangle}(M)$ with

$$\bar{q}^{\triangle} = \frac{1 + \frac{c_I\Delta_{II}^2}{c_{II}(c_I+c_{II}+\Delta_{II})^2}}{2 + \frac{c_I\Delta_{II}^2}{c_{II}(c_I+c_{II}+\Delta_{II})^2}}, \qquad r^{\triangle}(m) = \begin{cases} \text{risky}, & m \le \bar{q}^{\triangle}, \\ \text{safe}, & m > \bar{q}^{\triangle}, \end{cases}$$

from the first-order condition

$$0 = \frac{2c_I c_{II}}{c_I + c_{II}} \bar{q}^{\triangle} - \frac{2c_I \left((c_{II} + \Delta_{II})^2 + c_I c_{II} \right)}{(c_I + c_{II} + \Delta_{II})^2} \left(1 - \bar{q}^{\triangle}\right)$$

In the case of no reference dependence, $\Delta_{II} = 0$, the optimal threshold is always $\bar{q}^{\circ} = \frac{1}{2}$, no matter the relative costs. This stands in contrast to the threshold $\bar{p}^* = \frac{c_I}{c_I + c_{II}}$ of the machine taking the decision directly, which depends on the cost and is not generally the same as $\bar{q}^{\circ}$.

A.4 Equilibrium Action with Recommendation Dependence

We now solve for the action taken by the decision-maker given the optimal threshold. Plugging the solution back into into $a^{\bar{q}}$ yields the equilibrium action $A^{\triangle} = a^{\triangle}(H, R^{\triangle})$ with

$$a^{\triangle}(h, r) = \begin{cases} \text{risky}, & h \leq \bar{h}^{\triangle}(r), \\ \text{safe}, & h > \bar{h}^{\triangle}(r), \end{cases}$$

with recommendation-dependent thresholds

$$\bar{h}^{\triangle}(r) = \frac{c_I}{2\left(c_I + c_{II} + \Delta_{II}\right) + \frac{c_I \Delta_{II}^2}{c_{II}(c_I + c_{II} + \Delta_{II})}} + \frac{c_{II} + \Delta_{II} - \frac{\left(\frac{1}{c_I} + \frac{\Delta_{II}^2}{c_{II}(c_I + c_{II} + \Delta_{II})^2}\right)\left(c_{II}(c_{II} + \Delta_{II}) - c_I^2\right)}{2\left(\frac{c_I + c_{II} + \Delta_{II}}{c_I}\right) + \frac{\Delta_{II}^2}{c_{II}(c_I + c_{II} + \Delta_{II})}}}{c_I + c_{II}} \mathbb{1}(r = \text{risky})$$

obtained from combining the results in Section A.2 and Section A.3.

A.5 Strategic Non-Recommendation for the Example

Here we given a more complete development of the optimal thresholds $\bar{q}^{\text{low}}$ and $\bar{q}^{\text{high}}$ in the example, where recommendation dependence is considered when designing r.

We begin by arguing that at the optimum $\bar{q}^{\text{high}} = 2\bar{q}^{\text{low}}$. Consider the problem conditioned on $M \in [0, \bar{q}^{\text{high}}]$, which does not involve any recommendation dependence (where we still assume that $\Delta_I = 0$) since the machine gives a recommendation $R \neq$ safe in this regime. Setting the recommendation threshold for $\bar{q}^{\text{low}}$ optimally becomes a scaled version of the recommendation problem given in Section A.3 with $\Delta_{II} = 0$. By the same argument given there, the optimal threshold maximizes information by cutting the area into two equally-sized parts, $\bar{q}^{\text{low}} = \frac{\bar{q}^{\text{high}}}{2}$.

Defining both thresholds in terms of a single parameter x by $\bar{q}^{\text{low}} = x, \bar{q}^{\text{high}} = 2x$ simplifies threshold optimization to a quadratic optimization problem similar to the one considered in Section A.3,

$$
\begin{aligned}
\bar{q}^{\text{low}} = \arg\min_{x\in[0,1]} & \frac{c_I}{2}\left(\left[(1-2x)\frac{c_{II}+\Delta_{II}}{c_I+c_{II}+\Delta_{II}}\right]^2 + 2\left[(x)\frac{c_{II}}{c_I+c_{II}}\right]^2\right) \\
& + \frac{c_{II}}{2}\left(\left[(1-2x)\frac{c_I}{c_I+c_{II}+\Delta_{II}}\right]^2 + 2\left[(x)\frac{c_I}{c_I+c_{II}}\right]^2\right) \\
= & \frac{c_I\left((c_{II}+\Delta_{II})^2+c_Ic_{II}\right)}{2(c_I+c_{II}+\Delta_{II})^2}(1-2x)^2 + \frac{c_Ic_{II}}{c_I+c_{II}}x^2.
\end{aligned}
$$

We can check for interior minima to this quadratic program by taking a first order condition on the unconstrained problem,

$$
\begin{aligned}
& & 0 = -2\left[\frac{c_I\left((c_{II}+\Delta_{II})^2+c_Ic_{II}\right)}{(c_I+c_{II}+\Delta_{II})^2}(1-2x)\right] + 2\frac{c_Ic_{II}}{c_I+c_{II}} \\
\Rightarrow & \quad \frac{c_I\left((c_{II}+\Delta_{II})^2+c_Ic_{II}\right)}{(c_I+c_{II}+\Delta_{II})^2} = & \left(2\frac{c_I\left((c_{II}+\Delta_{II})^2+c_Ic_{II}\right)}{(c_I+c_{II}+\Delta_{II})^2} + \frac{c_Ic_{II}}{c_I+c_{II}}\right)x \\
\Rightarrow & & x = \frac{1}{2+\frac{c_{II}(c_I+c_{II}+\Delta_{II})^2}{(c_I+c_{II})((c_{II}+\Delta_{II})^2+c_Ic_{II})}}.
\end{aligned}
$$

As this minimum is always in $[0,1]$ and the problem is quadratic, we have found the global loss minimizer. Thus, the optimal choice of thresholds is obtained as

$$
\bar{q}^{\text{low}} = \frac{1}{2+\frac{c_{II}(c_I+c_{II}+\Delta_{II})^2}{(c_I+c_{II})((c_{II}+\Delta_{II})^2+c_Ic_{II})}}, \qquad \bar{q}^{\text{high}} = \frac{1}{1+\frac{c_{II}(c_I+c_{II}+\Delta_{II})^2}{2(c_I+c_{II})((c_{II}+\Delta_{II})^2+c_Ic_{II})}}.
$$

B Proofs for General Solution

Proof of Remark 1. An example is given in Section 3. □

Proof of Remark 2. For the case where not providing a recommendation can be better, consider the case where Δ_I, Δ_{II} are very high and the private machine information is substantially less helpful than the private information available to the decision-maker. In this case, with a recommendation, the decision-maker follows the recommendation closely to take a decision that is only weakly correlated with the optimal decision. Without a recommendation, on the other hand, the decision-maker takes a decision that better tracks the optimal decision.

If $\Delta_I = 0$ then always recommending $R =$ risky ($\bar{q}_x = 1$) does not distort decisions relative to the case of not providing a recommendation. An optimal choice can only (weakly) improve loss. □

Proof of Proposition 1. In the notation introduced after Proposition 2,

$$\begin{aligned} \mathrm{P}_x(A^\triangle = R|R = \text{risky}) &= \mathrm{P}_x(A^\triangle = \text{risky}|R = \text{risky}) \\ &= \mathrm{P}_x\left(P^{\text{risky}} \leq \frac{c_I + \Delta_I}{c_I + c_{II} + \Delta_I} \middle| Q \leq \bar{q}_x\right) \end{aligned}$$

where P^{risky} is unaffected by Δ_I and $\frac{c_I+\Delta_I}{c_I+c_{II}+\Delta_I}$ is monotonically increasing in Δ_I, which means that $\mathrm{P}_x(A^\triangle = R|R = \text{risky})$ can not decrease as Δ_I increases. The result for $\mathrm{P}_x(A^\triangle = R|R = \text{safe})$ follows simultaneously. □

Proof of Proposition 2. For every fixed threshold $\bar{q}_x$, we find that the expected loss

$$\begin{aligned} &\bar{\ell}_x(\bar{q}_x, \Delta_I, \Delta_{II}) = \mathrm{E}_x[\ell(Y, A^\triangle)] \\ &= \mathrm{E}_x\left[c_{II} P^{\text{risky}} \mathbb{1}(P^{\text{risky}} \leq \bar{p}^{\text{risky}}) + c_I(1-P^{\text{risky}})\mathbb{1}(P^{\text{risky}} > \bar{p}^{\text{risky}}) \middle| Q \leq \bar{q}_x\right] \ \mathrm{P}(Q \leq \bar{q}_x) \\ &\quad + \mathrm{E}_x\left[c_{II} P^{\text{safe}} \mathbb{1}(P^{\text{safe}} \leq \bar{p}^{\text{safe}}) + c_I(1-P^{\text{safe}})\mathbb{1}(P^{\text{safe}} > \bar{p}^{\text{safe}}) \middle| Q > \bar{q}_x\right] \ \mathrm{P}_x(Q{>}\bar{q}_x) \\ &= c_{II} \int_{p=0}^{\frac{c_I+\Delta_I}{c_I+c_{II}+\Delta_I}} p \ \mathrm{P}_x(Q{\leq}\bar{q}_x) \ f_x^{\text{risky}}(p;\bar{q}_x)\,\mathrm{d}p + c_I \int_{p=\frac{c_I+\Delta_I}{c_I+c_{II}+\Delta_I}}^{1} (1-p) \ \mathrm{P}_x(Q{\leq}\bar{q}_x) \ f_x^{\text{risky}}(p;\bar{q}_x)\,\mathrm{d}p \\ &\quad + c_{II} \int_{p=0}^{\frac{c_I}{c_I+c_{II}+\Delta_{II}}} p \ \mathrm{P}_x(Q{>}\bar{q}_x) \ f_x^{\text{safe}}(p;\bar{q}_x)\,\mathrm{d}p + c_I \int_{p=\frac{c_I}{c_I+c_{II}+\Delta_{II}}}^{1} (1-p) \ \mathrm{P}_x(Q{>}\bar{q}_x) \ f_x^{\text{safe}}(p;\bar{q}_x)\,\mathrm{d}p \end{aligned}$$

given the optimal action $A^\triangle$ converges to the risk

$$\begin{aligned} \bar{\ell}_x^m(\bar{q}_x) &= c_{II} \int_{p=0}^{1} p \ \mathrm{P}_x(Q{\leq}\bar{q}_x) \ f_x^{\text{risky}}(p;\bar{q}_x)\,\mathrm{d}p + c_I \int_{p=0}^{1} (1-p) \ \mathrm{P}_x(Q{>}\bar{q}_x) \ f_x^{\text{safe}}(p;\bar{q}_x)\,\mathrm{d}p \\ &= c_{II}\,\mathrm{P}_x(Y{=}\text{bad}|Q{\leq}\bar{q}_x) \ \mathrm{P}_x(Q{\leq}\bar{q}_x) + c_I\,\mathrm{P}_x(Y{=}\text{good}|Q{>}\bar{q}_x) \ \mathrm{P}_x(Q{>}\bar{q}_x) = \mathrm{E}_x[\ell(Y,R)] \end{aligned}$$

of directly implementing the recommendation. Furthermore,

$$\begin{aligned} &\sup_{\bar{q}_x} \left|\bar{\ell}_x(\bar{q}_x, \Delta_I, \Delta_{II}) - \bar{\ell}_x^m(\bar{q}_x)\right| \\ &= \sup_{\bar{q}_x} \left| \int_{p=\frac{c_I+\Delta_I}{c_I+c_{II}+\Delta_I}}^{1} (c_I(1-p) - c_{II}p) \ \mathrm{P}_x(Q{\leq}\bar{q}_x) \ f_x^{\text{risky}}(p;\bar{q}_x)\,\mathrm{d}p \right. \\ &\qquad \left. - \int_{p=0}^{\frac{c_I}{c_I+c_{II}+\Delta_{II}}} (c_I(1-p) - c_{II}p) \ \mathrm{P}_x(Q{\leq}\bar{q}_x) \ f_x^{\text{safe}}(p;\bar{q}_x)\,\mathrm{d}p \right| \\ &\leq \left(\frac{c_{II}}{c_I{+}c_{II}{+}\Delta_I} + \frac{c_I}{c_I{+}c_{II}{+}\Delta_{II}}\right)(c_I + c_{II}) \underbrace{\sup_{\bar{q}_x,p}\left(f_x^{\text{risky}}(p;\bar{q}_x) + f_x^{\text{safe}}(p;\bar{q}_x)\right)}_{<\infty} \to 0 \end{aligned}$$

as $\Delta_I \to \infty$ and $\Delta_{II} \to \infty$. Since $\bar{\ell}_x(\bar{q}_x, \Delta_I, \Delta_{II})$ converges to $\bar{\ell}_x^m(\bar{q}_x)$ uniformly in $\bar{q}_x$, and $\bar{\ell}_x^m(\bar{q}_x)$

is continuous with unique maximizer $\bar{p}^*$, we have that

$$\bar{q}_x^{\triangle}(\Delta_I, \Delta_{II}) = \arg\min_{\bar{q}_x} \bar{\ell}_x(\bar{q}_x, \Delta_I, \Delta_{II}) \to \arg\min_{\bar{q}_x} \bar{\ell}_x^m(\bar{q}_x) = \bar{p}^*$$

as $\Delta_I \to \infty$ and $\Delta_{II} \to \infty$. $\square$

Proof of Proposition 3. We note that for $X = x$ the expected loss from a recommendation $R = r_x(M)$ based on threshold $\bar{q}_x$ and optimal reference-dependent response $A^{\triangle} = a_x^{\triangle}(H, R)$ is

$$\begin{aligned}
&\bar{\ell}_x(\bar{q}_x, \Delta_I, \Delta_{II}) = \mathrm{E}_x[\ell(Y, A^{\triangle})] \\
&= \mathrm{E}_x\left[c_{II} P^{\text{risky}} \mathbb{1}(P^{\text{risky}} \leq \bar{p}^{\text{risky}}) + c_I(1 - P^{\text{risky}}) \mathbb{1}(P^{\text{risky}} > \bar{p}^{\text{risky}}) \middle| Q \leq \bar{q}_x\right] \ \mathrm{P}(Q \leq \bar{q}_x) \\
&\quad + \mathrm{E}_x\left[c_{II} P^{\text{safe}} \mathbb{1}(P^{\text{safe}} \leq \bar{p}^{\text{safe}}) + c_I(1 - P^{\text{safe}}) \mathbb{1}(P^{\text{safe}} > \bar{p}^{\text{safe}}) \middle| Q > \bar{q}_x\right] \ \mathrm{P}_x(Q > \bar{q}_x) \\
&= c_{II} \int_{p=0}^{\frac{c_I + \Delta_I}{c_I + c_{II} + \Delta_I}} p \ \mathrm{P}_x(Q \leq \bar{q}_x) \ f_x^{\text{risky}}(p; \bar{q}_x) \, \mathrm{d}p + c_I \int_{p=\frac{c_I + \Delta_I}{c_I + c_{II} + \Delta_I}}^{1} (1 - p) \ \mathrm{P}_x(Q \leq \bar{q}_x) \ f_x^{\text{risky}}(p; \bar{q}_x) \, \mathrm{d}p \\
&\quad + c_{II} \int_{p=0}^{\frac{c_I}{c_I + c_{II} + \Delta_{II}}} p \ \mathrm{P}_x(Q > \bar{q}_x) \ f_x^{\text{safe}}(p; \bar{q}_x) \, \mathrm{d}p + c_I \int_{p=\frac{c_I}{c_I + c_{II} + \Delta_{II}}}^{1} (1 - p) \ \mathrm{P}_x(Q > \bar{q}_x) \ f_x^{\text{safe}}(p; \bar{q}_x) \, \mathrm{d}p.
\end{aligned}$$

At a loss-minimizing choice $\bar{q}_x = \bar{q}_x^{\triangle}(\Delta_I, \Delta_{II})$, we have the first order condition

$$\frac{\partial}{\partial \bar{q}_x} \bar{\ell}_x(\bar{q}_x, \Delta_I, \Delta_{II})|_{\bar{q}_x = \bar{q}_x^{\triangle}(\Delta_I, \Delta_{II})} = 0.$$

We now consider the loss-minimizing choice $\bar{q}_x = \bar{q}_x^{\triangle}(\Delta_I, \Delta_{II})$ as Δ_I, Δ_{II} change infinitesimally. Since the first-order condition holds along all, say, Δ_I,

$$\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}\Delta_I} \left(\frac{\partial}{\partial \bar{q}_x} \bar{\ell}_x(\bar{q}_x, \Delta_I, \Delta_{II})|_{\bar{q}_x = \bar{q}_x^{\triangle}(\Delta_I, \Delta_{II})} \right) \\
&= \frac{\partial^2}{\partial \Delta_I \partial \bar{q}_x} \bar{\ell}_x(\bar{q}_x, \Delta_I, \Delta_{II})|_{\bar{q}_x = \bar{q}_x^{\triangle}(\Delta_I, \Delta_{II})} + \frac{\mathrm{d}\bar{q}_x^{\triangle}(\Delta_I, \Delta_{II})}{\mathrm{d}\Delta_I} \left(\frac{\partial^2}{\partial \bar{q}_x^2} \bar{\ell}_x(\bar{q}_x, \Delta_I, \Delta_{II})|_{\bar{q}_x = \bar{q}_x^{\triangle}(\Delta_I, \Delta_{II})} \right) \\
&= \overbrace{\frac{\partial \bar{p}_x^{\text{risky}}}{\partial \Delta_I} \left(c_{II} \, \bar{p}_x^{\text{risky}} - c_I \, (1 - \bar{p}_x^{\text{risky}})\right)}^{= \frac{\Delta_I \, c_{II}^2}{(c_I + c_{II} + \Delta_I)^3} \geq 0} \overbrace{\frac{\partial}{\partial \bar{q}_x} \left(\mathrm{P}_x(Q \leq \bar{q}_x) \ f_x^{\text{risky}}(\bar{p}_x^{\text{risky}}; \bar{q}_x) \right)}^{\geq 0} \\
&\quad + \frac{\mathrm{d}\bar{q}_x^{\triangle}(\Delta_I, \Delta_{II})}{\mathrm{d}\Delta_I} \underbrace{\left(\frac{\partial^2}{\partial \bar{q}_x^2} \mathrm{E}_x[\ell(Y, A^{\triangle})]|_{\bar{q}_x = \bar{q}_x^{\triangle}(\Delta_I, \Delta_{II})} \right)}_{>0}.
\end{aligned}$$

As a consequence, $\frac{\mathrm{d}\bar{q}_x^{\triangle}(\Delta_I, \Delta_{II})}{\mathrm{d}\Delta_I} \leq 0$, and analogously for $\frac{\mathrm{d}\bar{q}_x^{\triangle}(\Delta_I, \Delta_{II})}{\mathrm{d}\Delta_{II}}$ where $\frac{\partial \bar{p}_x^{\text{safe}}}{\partial \Delta_{II}} \leq 0$, where we rely on the implicit function theorem to argue that $\bar{q}_x^{\triangle}(\Delta_I, \Delta_{II})$ is continuously differentiable. $\square$

C Additional Proofs

Proof of Proposition 4. With an optimal choice of thresholds, by the envelope theorem (Roy's identity) the marginal effect on expected losses of a change in Δ_I, Δ_{II} is equal to the mechanical (partial) effect. Considering the case of Δ_I,

$$\frac{\mathrm{d}}{\mathrm{d}\Delta_I}(\mathrm{E}_x[\ell(Y, A^{\triangle})] - \mathrm{E}_x[\ell(Y, A^{\blacktriangle})]) = \frac{\partial}{\partial \Delta_I}(\mathrm{E}_x[\ell(Y, A^{\triangle})] - \mathrm{E}_x[\ell(Y, A^{\blacktriangle})])$$
$$= \underbrace{(c_{II}\, \bar{p}_x^{\text{risky}} - c_I\, (1 - \bar{p}_x^{\text{risky}}))}_{\geq 0} \underbrace{\left(\mathrm{P}_x(Q \leq \bar{q}_x)\, f_x^{\text{risky}}(\bar{p}_x^{\text{risky}}; \bar{q}_x) - \mathrm{P}(Q \leq \bar{q}_x^{\text{low}})\, f_x^{\text{risky}}(\bar{p}_x^{\text{risky}}; \bar{q}_x^{\text{low}})\right)}_{\geq 0} \geq 0$$

similar to the proof of Proposition 3. □

Proof of Proposition 5. Writing out this reference-dependent loss with loss aversion for the specific loss functions, we find that

$$\begin{aligned}
\ell^{\mathrm{PT}}(y, r, a) &= \lambda[\ell(y, a) - \ell(y, r)]_+ - \ [\ell(y, a) - \ell(y, r)]_- \\
&= \ell(y, a) - \ell(y, r) + (\lambda - 1)[\ell(y, a) - \ell(y, r)]_+ \\
&= \ell(y, a) - \ell(y, r) + \begin{cases} (\lambda - 1)c_I, & y = \text{good}, a = \text{safe}, r = \text{risky}, \\ (\lambda - 1)c_{II}, & y = \text{bad}, a = \text{risky}, r = \text{safe}. \end{cases}
\end{aligned}$$

Since $\ell(y, r)$ is not affected by the decision-maker's choice, their preferences are as if they are minimizing expected loss with loss function

$$\ell^*(y, a, r) = \ell(y, a) + \begin{cases} (\lambda - 1)c_I, & y = \text{good}, a = \text{safe}, r = \text{risky}, \\ (\lambda - 1)c_{II}, & y = \text{bad}, a = \text{risky}, r = \text{safe}, \end{cases}$$

as in (4) with $\Delta_I = (\lambda - 1)c_I, \Delta_{II} = (\lambda - 1)c_{II}$. □