https://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–17
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Is Your Machine Better Than You? You May Never Know

Francis de Véricourt,[a] Huseyin Gurkan[a,*]

[a] European School of Management and Technology Berlin, 10178 Berlin, Germany
*Corresponding author
Contact: francis.devericourt@esmt.org, https://orcid.org/0000-0003-2832-0738 (FdV); huseyin.gurkan@esmt.org, https://orcid.org/0000-0002-3872-2776 (HG)

Received: May 17, 2022
Revised: December 8, 2022
Accepted: February 6, 2023
Published Online in Articles in Advance: May 25, 2023

https://doi.org/10.1287/mnsc.2023.4791

Abstract. Artificial intelligence systems are increasingly demonstrating their capacity to make better predictions than human experts. Yet recent studies suggest that professionals sometimes doubt the quality of these systems and overrule machine-based prescriptions. This paper explores the extent to which a decision maker (DM) supervising a machine to make high-stakes decisions can properly assess whether the machine produces better recommendations. To that end, we study a setup in which a machine performs repeated decision tasks (e.g., whether to perform a biopsy) under the DM's supervision. Because stakes are high, the DM primarily focuses on making the best choice for the task at hand. Nonetheless, as the DM observes the correctness of the machine's prescriptions across tasks, the DM updates the DM's belief about the machine. However, the DM is subject to a so-called verification bias such that the DM verifies the machine's correctness and updates the DM's belief accordingly only if the DM ultimately decides to act on the task. In this setup, we characterize the evolution of the DM's belief and overruling decisions over time. We identify situations under which the DM hesitates forever whether the machine is better; that is, the DM never fully ignores but regularly overrules it. Moreover, the DM sometimes wrongly believes with positive probability that the machine is better. We fully characterize the conditions under which these learning failures occur and explore how mistrusting the machine affects them. These findings provide a novel explanation for human–machine complementarity and suggest guidelines on the decision to fully adopt or reject a machine.

History: Accepted by Elena Katok, special issue on the human–algorithm connection.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2023.4791.

Keywords: machine accuracy • decision making • human in the loop • algorithm aversion • dynamic learning

1. Introduction

The adoption of machine learning (ML) algorithms is revolutionizing the delivery of products and services (McKendrick 2021), especially in domains that require human expertise, such as the medical and judiciary sectors. Indeed, artificial intelligence tools demonstrate a capability to produce higher quality predictions than human judgment for many decision tasks (Grady 2019, Reardon 2019). The deployment of these tools in practice, however, is limited (Wiens et al. 2019) and challenged by the tendency of decision makers to override—sometimes wrongly—algorithmic prescriptions. For instance, Sun et al. (2021) find in warehouse operations that employees significantly deviate from the recommendations of an algorithm. Lebovitz et al. (2022) also report how a team of radiologists in a large U.S.-based hospital abandoned different ML algorithms after using them for several months.

This tendency to override algorithms is typically attributed to an intrinsic mistrust of machine-based predictions, often referred to as an algorithm aversion (Dietvorst et al. 2015, Gaube et al. 2021). This bias, however, may not be the sole reason for inappropriately and systematically overriding an algorithm. Indeed, the very context in which a human decision maker (DM) works can also prevent the DM from learning whether a machine produces better prescriptions.

In this paper, we explore the conditions under which making high-stakes decisions hampers the DM's ability to properly learn whether a machine is superior to human expertise. Importantly, we characterize the nature of the inappropriate overriding decisions to which these learning failures give rise without relying on any mistrust bias. To that end, we analyze a setup in which a DM performs repeated decision tasks using the prescriptions of a machine. Each task consists of deciding whether to take a specific action. This corresponds, for instance, to deciding on a biopsy in a medical context. To make this choice, the machine produces a recommendation that the DM may overrule based on the DM's own expertise. Crucially, the DM is uncertain about whether the machine makes better or worse decisions than the DM does, but as the DM

verifies the correctness of the different machine's prescriptions, the DM forms a belief about the machine's true accuracy.

Our focus is, thus, on the DM's learning behavior once the algorithm is deployed, that is, after it has been properly trained and evaluated on representative data sets (see Kubat 2017) and possibly shown better-than-human accuracy levels. These data sets, however, never fully capture the ground truth, and the issue of empirical generalizability remains (see, e.g., Lebovitz et al. 2021). Hence, an expert may continue to observe and adjust the expert's belief about the machine after adopting it. Yet, because the machine is deployed and makes prescriptions with real consequences, learning can be impaired in ways that do not exist during the training phase of the algorithm.

In particular, we consider situations in which the DM observes the correctness of the machine's prediction and updates the DM's belief accordingly only if the action is actually taken (e.g., when a biopsy is performed). In other words, the DM is subject to the so-called verification bias (see, e.g., Pepe 2003) such that the accuracy parameters of a diagnostic test are learned only when a test result is verified by follow-up work (e.g., a biopsy reveals the presence or absence of the disease). This limitation can also stem from a form of salience bias or inattentional blindness (Taylor and Thompson 1982, Bordalo et al. 2012, Tiefenbeck et al. 2018), which are especially prevalent for high-stakes decisions (see Lee et al. 2018). In this context, the DM focuses the DM's limited attention on making the decisions at hand but is triggered to reassess the machine's quality by the salient observation of a verified success or failure. (See Camacho et al. 2011 for an example of similar salient effects in the context of new drug prescriptions.)

Further, the DM only decides what is best for the task at hand and, thus, never acts for the purpose of verifying the machine's accuracy. In this sense, the DM's decisions are exploration-free. This restriction may be for legal or ethical concerns, which are often warranted when the stakes are high as in the medical and judiciary sectors (Bastani et al. 2021b).

In this paper, we mainly examine the case in which the machine and the DM are substitutes in that the DM's accuracy is either better or worse than the machine's. We focus on substitution for two reasons. First, our goal is to study inappropriate overriding decisions without relying on algorithm aversion, the studies of which assume substitution (Dietvorst et al. 2015, Sun et al. 2021). Second and more importantly, we seek to determine if a complementarity between the DM and the machine might emerge from the DM's inability to learn the nature of the machine. Assuming substitution enables us to disentangle this learning effect from an intrinsic complementarity between the DM and the machine. Nonetheless, we also explore situations in which the machine and DM complement one another (see Section 8).

Our approach, thus, consists of analytically studying the evolution of the DM's belief and overruling decisions over time. This enables identification of situations in which the DM properly learns whether the machine makes better predictions than the DM does. The asymptotic behavior of the DM's belief further characterizes the different ways in which the DM fails to learn the true nature of the machine.

Following this approach, we find that the DM always properly learns whether the machine is better or worse in the absence of human–machine interactions, that is, when the machine's prescriptions never influence the DM's decisions. Indeed, in this case, the DM verifies the machine independently of its prediction.[1] Hence, inappropriate overriding decisions may occur only if the machine has some influence on the DM's choices.

When this influence is maximal, that is, the machine's prescriptions fully determine the DM's choices, we find that the DM's ability to learn depends on the DM's prior about the task. Specifically, the DM properly learns that the machine is better (respectively, worse), if the prior probability that an action is required for the task at hand is above (below) a certain threshold. Hence, the DM can end up believing that the machine makes worse predictions than the DM does even though the machine is actually better. This occurs when the action is not too frequently required for the tasks. Conversely, the DM learns that the machine is better even though it is actually worse when the action is frequently required.

In these two benchmarks, the DM's belief about the machine has no effect on the DM's choices. This contrasts with our main setup, in which the DM's decision to act and, hence, the DM's ability to learn, are endogenously determined by the DM's current belief about the machine's quality. Specifically, in this setting, the DM overrules the machine when the DM's judgment contradicts the machine's prescription and the DM sufficiently believes that the machine is worse.

When this is the case, we again find that the prior about the task determines when the DM fails to learn. However, the DM's overriding decisions fundamentally change the nature of mislearning. Indeed, when the machine is actually better than the DM and the prior about the task is low, the DM's belief always oscillates over time. In other words, the DM permanently remains unsure about whether the machine is better or not and constantly alternates between following and overriding its prescriptions. Further, and perhaps more interestingly, the DM sometimes treats the machine as if its prediction complements the DM's own judgment, whereas, in fact, the two are full substitutes. In contrast, when the machine is worse and the prior about the task is high, the belief converges to a Bernoulli random variable: the

DM properly learns that the machine is worse with a given probability but wrongly learns that it is better with the remaining probability. In other words, the DM randomly ends up incorrectly believing that the machine is better.

Taken together, these results identify two different forms of mislearning—persistent hesitation and random inference—that can occur when a DM works with a machine to make high-stakes decisions. These findings also highlight the key role that the DM's prior about the task plays in the DM's ability to learn the true nature of the machine. Additionally, they uncover a novel rationale—the uncertainty about the machine's true performance—for why human experts may coproduce their decisions with a machine. These mislearning behaviors do not depend on the DM's initial belief about the machine and, thus, hold even when the DM sufficiently believes in the machine's performance to deploy it in the first place.

These results further suggest guidelines on the decision to fully adopt or reject a machine after it has been deployed. The question is whether the machine should make all the decisions henceforth or be abandoned for good at some point after working with it. Our results indicate that the longer the DM believes the machine is worse, the more likely the DM is correct in the DM's assessment and, hence, should abandon the machine. The same is not true, however, if the DM increasingly believes that the machine is better. In this case, our findings suggest relying on multiple DMs. If a consensus exists among the team that the machine is better, then the larger the team is, the more likely it is that the machine should be adopted (see Section 6).

Importantly, the mislearning behaviors we characterize in this paper do not stem from an intrinsic algorithm aversion, but rather from certain contexts in which DMs make high-stakes decisions as captured with the verification bias and the exploration-free condition. Yet a DM who faces situations such as these may also be subject to mistrust biases against the machine, which can interact with our findings.

Indeed, mistrusting the machine affects the DM's ability to learn in at least two ways. First, the DM may downplay the machine's prescription when deciding to act (consistently with the decision-making literature; see, e.g., Soll and Mannes 2011), which alters the DM's ability to observe the correctness of the machine's predictions. Second, and in line with the algorithm aversion reported by Dietvorst et al. (2015), the DM's belief in the machine may disproportionately drop upon observing a machine's prediction error. We explore how these effects interact with our results (see Section 7) and find that our results hold in the former case but not in the latter. When mistrust introduces a negativity bias in the DM's learning process, the DM does not always properly learn that the machine is better if the prior is sufficiently high as in the main setup. Instead, the DM can wrongly learn with a positive probability that the machine is worse. In this sense, algorithm aversion sometimes interacts with our setting to randomize the DM's ability to learn.

Finally, our results are robust to a partial relaxation of the verification bias, which is legitimate, for instance, when the bias stems from the DM's limited attention. In this context, the DM also learns from unverified cases. Our results continue to hold as long as unverified cases are sufficiently less salient than verified ones, for which the true state of the world is revealed.

After reviewing the literature in Section 2, we present the model in Section 3. In Section 4, we analyze the no-interaction and no-overriding benchmarks and then focus on the main setup in Section 5. We highlight the implications of our findings in Section 6 and study the effects of mistrust biases on our findings in Section 7. Further, we explore the settings in which the machine and the DM complement one another in Section 8, and the verification bias is partially relaxed in Section 9. Finally, we discuss future research directions in the conclusion.

2. Literature Review

Our study is related to the recent and growing literature on the interaction between human decision makers and data-driven algorithms. This research explores the extent to which coproduction of decisions by a machine and a DM may improve performance. For instance, Boyacı et al. (2023) demonstrate in a rational inattention framework that human–machine interaction improves the overall accuracy of decisions but sometimes at the cost of higher cognitive effort (see Boyacı et al. 2023 for additional references on formal models of machine–human interactions). Machine learning algorithms are also proposed to provide interpretable cues to help decision makers improve their decisions (see Bastani et al. 2021a, for instance). This stream of research further explores how to use human judgment to train or improve an algorithm (Van Donselaar et al. 2010, Cowgill 2019, Ibrahim et al. 2021).

We contribute to this literature by providing a novel rationale for why a DM may treat the machine's prescriptions as a complement to the DM's judgment. In fact, this stream of research typically assumes that the machine's accuracy is known and complements the DM's judgment. In contrast, the DM and the machine are substitutes, and the machine's accuracy is unknown in our setting.

In this sense, our study is closely related to the literature on overriding decisions and, more generally, trust in algorithmic prescriptions. In particular, Lebovitz et al. (2021) document over several months how a team of radiologists lost trust in the quality of a machine learning algorithm that helped analyze medical images. Dietvorst

et al. (2015) also find in an experimental setup that their participants overrode a machine's prescriptions even after seeing that the machine's algorithm performed better than the human did on average. This tendency to wrongly override machine-based prescriptions is further supported by empirical evidence in the field. For instance, Sun et al. (2021) observe that packing workers at the warehouses of the Alibaba Group regularly deviated from algorithmic prescriptions, which reduced operational efficiency. Several approaches are explored to reduce deviations such as these either with field experiments (Sun et al. 2021) or in the laboratory (Dietvorst et al. 2018).

In contrast to this stream of papers, our study proposes an alternative explanation for inappropriately overriding decisions such as these, which mostly stems from the context in which the decisions are made. Specifically, we trace these errors to four fundamentals (exploration-free, verification bias, informativeness, and substitution), which capture some essential features of high-stakes decision making using machine-based predictions.

Recent studies also suggest that humans follow the principles of Bayesian inference when observing the correctness of machine-based decisions. For instance, Wang et al. (2018) and Guo et al. (2020) analyze in an experimental setup how observers dynamically update their trust in the machine as they observe the failures and successes of its predictions (without overriding the machine as in the benchmark of Section 4.2). These studies find that assuming Bayesian observers can explain the empirical level of human trust in the machine over time. The key difference with our setup, however, is that the DM is not subject to verification bias and, thus, always observes the correctness of the machine's prediction in their settings.

Verification bias is a form of selection bias that was first introduced by Ransohoff and Feinstein (1978) to describe situations in which the accuracy of a diagnostic test is learned only with the verified cases, that is, when follow-up actions are taken to confirm a test result. This bias toward verified cases permeates the medical field (e.g., Greenes and Begg 1985, Bates et al. 1993, Petscavage et al. 2011, Whiting et al. 2013, Hujoel et al. 2021) and is found in studies evaluating ML algorithms for medical applications (see, e.g., Tschandl et al. 2019). Most of this research focuses on developing estimators of accuracy based on the maximum likelihood to correct the bias. The conditions required to avoid this bias in the frequentist literature, however, do not hold in our setup.[2]

Learning with selective observations as in verification bias can also stem from a form of salience bias or inattentional blindness (Taylor and Thompson 1982, Bordalo et al. 2012, Tiefenbeck et al. 2018). The behavioral science literature studies biases such as these (see, e.g., Kahneman 1973, chapter 7, for the effects of focused attention on information filtering), which are due to the DM's limited cognitive capacity (Simon 1955). In this sense, our study also contributes to the growing stream of operations and economics literature, which deviates from standard Bayesian learning to account for limited cognitive capacity (see, e.g., Allon et al. 2021, Boyacı et al. 2023 and the references therein). In particular, our setup is consistent with the notion of selective Bayesian updating (see the seminal work of Schwartzstein 2014) in that the DM only selects the actual success or failure of the machine to update the DM's belief about the machine's type.

Finally, our work is related to the vast literature on learning problems, which are extensively studied in management science and operations management. For instance, studies consider price experimentation to learn demand curves by focusing on the trade-offs between learning and earning and design heuristic policies achieving good regret performance (Besbes and Zeevi 2009, Boyacı and Özer 2010, Cheung et al. 2017, Keskin and Birge 2019). In this stream of papers, the DM experiments (explores) with different prices in the beginning of the time horizon to earn (exploit) more in the remaining periods. Because of this ability to explore, the DM can, in principle, properly uncover the true demand curve in the limit. The objective of these papers is then to learn sufficiently fast so as to maximize profit. In contrast, we consider situations in which exploring is not possible. Thus, the DM optimizes within each period and mislearning may emerge in our setup.

In this sense, our approach resembles Harrison et al. (2012) who analyze myopic pricing policies (see section 4 in particular). In their setup, demand functions are the focus of learning, whereas we consider unknown accuracy parameters. Therefore, the type of incomplete learning that may occur differs radically in each setting. In particular, incomplete learning takes the form of confounding beliefs in Harrison et al. (2012) such that the myopic policy charges an uninformative price, which prevents Bayesian updating from producing a different posterior. As a result, the DM becomes stuck in the same belief over time. In contrast, mislearning can take the form of belief oscillation in our setup, which cannot occur in Harrison et al. (2012) per proposition 2.

Learning problems such as these are also extensively studied in economics (see, for instance, Acemoglu et al. 2011, Smith and Sørensen 2000, and references therein) with a particular focus on equilibrium learning dynamics shaped by multiple strategic agents. In this stream of research, Herrera and Hörner (2013) analyze a setup with short-lived myopic investors in which only investing decisions are observable. Although this may resemble our setup, their payoff, signal and learning structures differ, which yields a different type of mislearning. In particular, the belief converges to an interior point in their

setup (see propositions 1 and 4 in Herrera and Hörner 2013), whereas it may not converge in ours.

3. Model Description

We consider a DM who faces a series of independent decision tasks over a discrete-time infinite horizon. A machine further assists the DM by producing a recommendation about which decision to take for each task. The DM, however, does not know if the machine's accuracy is superior to the DM's own judgment. As the accuracy of the machine's predictions is revealed over time, the DM forms a belief about whether the DM should override the machine's prediction. Next, we introduce the single decision task problem that the DM performs in each period. We then consider the whole time horizon.

3.1. Single Decision Task

A task consists of deciding whether a specific action (e.g., a biopsy) is required. We denote as $\Theta \in \{\mathsf{A}, \mathsf{NA}\}$ the type of task such that the action is required if $\Theta = \mathsf{A}$ and is not required if $\Theta = \mathsf{NA}$. The DM does not know the task's type but has a prior belief $p \triangleq \mathbb{P}(\Theta = \mathsf{A})$ that the DM should act.

To perform this task, the DM applies the DM's expertise and elicits an imperfect signal $S^{\mathsf{H}} \in \{+, -\}$ such that $S^{\mathsf{H}} = +$ ($S^{\mathsf{H}} = -$) indicates that $\Theta = \mathsf{A}$ ($\Theta = \mathsf{NA}$). We denote the sensitivity (true positive rate) and specificity (true negative rate) of the signal by α^{H} and β^{H}, respectively. The DM is further assisted by a machine learning algorithm, which makes an independent prediction about type Θ. This prediction corresponds to a second signal, $S^{\mathsf{M}} \in \{+, -\}$, with sensitivity and specificity equal to $(\alpha^{\mathsf{M}}, \beta^{\mathsf{M}})$.

Importantly, the DM is uncertain about whether the machine's accuracy is better than the DM's own. Specifically, we denote the machine's type as $\Gamma \in \{\mathsf{B}, \mathsf{W}\}$. When $\Gamma = \mathsf{B}$ ($\Gamma = \mathsf{W}$), signal S^{M} is better (worse) than signal S^{H}, and the sensitivity and specificity of the signal are equal to $(\alpha^{\mathsf{B}}, \beta^{\mathsf{B}})$ $((\alpha^{\mathsf{W}}, \beta^{\mathsf{W}}))$. The machine is better (worse) in the sense that the DM never (always) overrules the machine when its type is perfectly known. This corresponds to the notion of substitution, which we introduce and formalize later in this section (see Equations (4) and (5)). To exclude degenerated cases, we further focus our analysis on situations in which $\alpha^{\mathsf{B}} > \alpha^{\mathsf{W}}$ and $\beta^{\mathsf{B}} > \beta^{\mathsf{W}}$.[3] This is only for the sake of clarity as all of our results extend to the more general case.

Probability $b \triangleq \mathbb{P}(\Gamma = \mathsf{B})$ denotes then the DM's belief that the machine outperforms the DM's ability to decide. In effect, these two types of machine induce two different probability measures $\mathbb{P}^{\mathsf{B}}\{\cdot\}$ and $\mathbb{P}^{\mathsf{W}}\{\cdot\}$ on the sample space of the machine's signals such that $\mathbb{P}^{\Gamma}(S^{\mathsf{M}} = +, \Theta = \mathsf{A}) = \alpha^{\Gamma} p$ and $\mathbb{P}^{\Gamma}(S^{\mathsf{M}} = -, \Theta = \mathsf{NA}) = \beta^{\Gamma} \overline{p}$ for $\Gamma \in \{\mathsf{B}, \mathsf{W}\}$ (with $\overline{x} = 1 - x$ for $x \in [0, 1]$).

Based on realizations s^{H} and s^{M} of signals S^{H} and S^{M}, respectively, and the DM's belief b about the machine, the DM updates the DM's prior p that an action is required using Bayes' rule. The corresponding posterior probability is, thus, $\mathbb{P}(\Theta = \mathsf{A} | S^{\mathsf{H}} = s^{\mathsf{H}}, S^{\mathsf{M}} = s^{\mathsf{M}}, b)$ (with a slight abuse of notation).[4]

The DM then decides to act if and only if the posterior is above a positive threshold r, that is, $\mathbb{P}(\Theta = \mathsf{A} | S^{\mathsf{H}} = s^{\mathsf{H}}, S^{\mathsf{M}} = s^{\mathsf{M}}, b) \geq r$; the DM does not act otherwise. This decision rule is optimal, for instance, when the DM seeks to maximize the expected value associated with correctly identifying the task's type. In this case, threshold r accounts for the false positive and false negative costs associated with the decision.[5]

3.1.1. Informativeness.

In the following, we assume that the signals produced by both the DM and the machine are informative in the sense that each signal provides sufficient information for the DM to decide. Formally, this corresponds to

$$\mathbb{P}(\Theta = \mathsf{A} | S^{\mathsf{H}} = +) \geq r \text{ and } \mathbb{P}(\Theta = \mathsf{A} | S^{\mathsf{H}} = -) < r, \tag{1}$$

$$\mathbb{P}^{\mathsf{B}}(\Theta = \mathsf{A} | S^{\mathsf{M}} = +) \geq r \text{ and } \mathbb{P}^{\mathsf{B}}(\Theta = \mathsf{A} | S^{\mathsf{M}} = -) < r, \tag{2}$$

$$\mathbb{P}^{\mathsf{W}}(\Theta = \mathsf{A} | S^{\mathsf{M}} = +) \geq r \text{ and } \mathbb{P}^{\mathsf{W}}(\Theta = \mathsf{A} | S^{\mathsf{M}} = -) < r. \tag{3}$$

In other words, the sole realization of either S^{H} or S^{M}, whether the machine is of type B or W, fully determines whether the posterior is larger than threshold r, that is, the DM takes the action. These conditions further imply that considering both signals S^{H} and S^{M} together is redundant when their realizations are aligned, that is, when $s^{\mathsf{H}} = s^{\mathsf{M}}$. One signal is then sufficient for the DM to decide because the DM acts if $S^{\mathsf{H}} = S^{\mathsf{M}} = +$ and does not act if $S^{\mathsf{H}} = S^{\mathsf{M}} = -$. If the realizations are misaligned with $s^{\mathsf{H}} \neq s^{\mathsf{M}}$, however, the DM and the machine may override one another. In this case, we consider situations in which the machine and the DM are full substitutes in the following sense.

3.1.2. Substitution.

We assume that a type B machine always overrides the DM's judgment, whereas the DM always overrides the prescription of a type W machine. Formally, this corresponds to

$$\mathbb{P}^{\mathsf{B}}(\Theta = \mathsf{A} | S^{\mathsf{H}} = +, S^{\mathsf{M}} = -) < r \text{ and}$$
$$\mathbb{P}^{\mathsf{B}}(\Theta = \mathsf{A} | S^{\mathsf{H}} = -, S^{\mathsf{M}} = +) \geq r, \tag{4}$$

$$\mathbb{P}^{\mathsf{W}}(\Theta = \mathsf{A} | S^{\mathsf{H}} = +, S^{\mathsf{M}} = -) \geq r \text{ and}$$
$$\mathbb{P}^{\mathsf{W}}(\Theta = \mathsf{A} | S^{\mathsf{H}} = -, S^{\mathsf{M}} = +) < r. \tag{5}$$

Thus, if the signals of the DM and a type B machine contradict one another, signal S^{M} alone determines whether the posterior probability is larger than the threshold (per

Equation (4)). Along with the informativeness property, this means that a type B machine always determines whether the DM should act, independently of the DM's judgment. In contrast, the DM decides alone and can ignore the prescription of a type W machine (per Equation (5)). Hence, if the machine's type is fully known, the DM and the machine never collaborate to make a decision. In this sense, the DM and the machine are substitutes for the task.

In essence, informativeness and substitution are conditions on the DM's posterior probability about the task's type, which, in turn, depends on the signals' sensitivities and specificities as well as prior p and threshold r.

3.2. Repeated Tasks and Learning

We now consider the situation in which the DM faces a series of decision tasks over a discrete time infinite horizon. Task types Θ_t, $t \in \mathbb{N}$, are independent and identically distributed with probability p. (In the following, we use subscript t to denote the parameters associated with the task of period t.) At the beginning of period $t > 0$, the DM's belief about the machine's type is given by b_{t-1}, where b_0 is the prior belief at the beginning of the time horizon. The DM then obtains signals S_t^{H}, S_t^{M} and decides whether to act.

3.2.1. Exploration-Free. In making this choice, the DM considers only the task at hand. More formally, the DM acts if $\mathbb{P}(\Theta_t = \mathsf{A} | S_t^{\mathsf{H}}, S_t^{\mathsf{M}}, b_{t-1}) \geq r$ and does nothing otherwise. In particular, the DM does not act for the sole purpose of uncovering the true task's type and, thus, learning the machine's. Instead, the DM decides what the DM thinks is best for the current task and is, thus, myopic with respect to learning the machine's type.

At the end of the period, the DM updates the DM's belief b_{t-1} to posterior b_t according to Bayes' rule if the DM observes type Θ_t.

3.2.2. Verification Bias. The DM, however, observes the task's type and updates the DM's belief accordingly if and only if an action is taken. Because decisions are exploration-free, the verification bias implies that the DM updates the DM's belief if and only if $\mathbb{P}(\Theta_t = \mathsf{A} | S_t^{\mathsf{H}}, S_t^{\mathsf{M}}, b_{t-1}) \geq r$, in which case we assume that the DM follows Bayes' rule. Thus, we have

$$b_t = \begin{cases} b_{t-1} & \text{if } \mathbb{P}(\Theta_t = \mathsf{A} \mid S_t^{\mathsf{H}} = s^{\mathsf{H}}, S_t^{\mathsf{M}} = s^{\mathsf{M}}, b_{t-1}) < r \\ \left[1 + \dfrac{\overline{b}_{t-1}}{b_{t-1}} \dfrac{\mathbb{P}^{\mathsf{W}}(S_t^{\mathsf{M}} = s^{\mathsf{M}} \mid \Theta_t = \theta)}{\mathbb{P}^{\mathsf{B}}(S_t^{\mathsf{M}} = s^{\mathsf{M}} \mid \Theta_t = \theta)}\right]^{-1} & \\ \qquad \text{if } \mathbb{P}(\Theta_t = \mathsf{A} \mid S_t^{\mathsf{H}} = s^{\mathsf{H}}, S_t^{\mathsf{M}} = s^{\mathsf{M}}, b_{t-1}) \geq r, \end{cases} \tag{6}$$

where $\theta \in \{\mathsf{A}, \mathsf{NA}\}$ is the observed value of Θ_t.

Equation (6) highlights two mechanisms by which the DM's belief about the machine's type is endogenously determined over time. The first corresponds to the Bayesian updating of prior b_{t-1} when the DM observes type Θ_t. The second corresponds to the DM's ability to verify type Θ_t in the first place, that is, whether posterior belief $\mathbb{P}(\Theta_t = \mathsf{A} | S_t^{\mathsf{H}}, S_t^{\mathsf{M}}, b_{t-1})$ is sufficiently large. This, in turn, depends on belief b_{t-1}. Equation (6) further implies that, when the DM acts, the DM increases (decreases) the DM's belief that the machine is better if the machine's prescription turns out to be correct (wrong).

This learning mechanism also corresponds to a selective Bayesian updating setup (Schwartzstein 2014) in which a DM focuses the DM's limited attention on making diagnostic decisions instead of evaluating the machine that assists the DM. The salient observation of an actual success or failure of the machine, however, redirects the DM's attention to reassess the DM's belief about the machine's type. This mechanism resembles the two-stage learning process of Allon et al. (2021), in which agents allocate their attention to different tasks (screening and belief updating in their setting) in each stage. (We relax this limitation on the DM's attention in Section 9.)

When the DM acts, the DM updates belief b_t in part based on signal S_t^{M}. The machine's type, however, determines the probability distribution, $\mathbb{P}^{\mathsf{B}}\{\cdot\}$ or $\mathbb{P}^{\mathsf{W}}\{\cdot\}$, of this signal. Hence, belief $(b_t)_{t \in \mathbb{N}}$ can follow two different stochastic processes depending on machine type Γ. The asymptotic behavior of belief b_t, thus, captures the DM's ability to learn whether the machine makes better predictions. Indeed, the DM properly learns the machine's type if the DM's belief converges over time to one ($b_t \overset{\text{a.s.}}{\longrightarrow} 1$) when the machine is better ($\Gamma = \mathsf{B}$) and converges to zero ($b_t \overset{\text{a.s.}}{\longrightarrow} 0$) when the machine is worse ($\Gamma = \mathsf{W}$). (Notation $\overset{\text{a.s.}}{\longrightarrow}$ indicates almost-sure convergence.) In contrast, the DM mislearns the machine's type when $b_t \overset{\text{a.s.}}{\longrightarrow} 0$ (1) and $\Gamma = \mathsf{B}$ ($\Gamma = \mathsf{W}$). Learning may even be inconclusive when belief b_t converges to an interior point in (0, 1) or oscillates. More formally, a stochastic process Y_t is said to be oscillating and recurrent if recurrent interval $\mathcal{I}$ exists such that, for any $\tau > 0$, $\mathbb{P}(Y_t \in \mathcal{I}$ for some $t > \tau | Y_\tau \in \mathcal{I}) = 1$ (see definition 8.1 in Gut 2009 for instance).

Our objective, therefore, is to study the asymptotic behavior of b_t and characterize the resulting learning behavior of the DM.

4. Benchmarks

We first study two settings in which the DM does not account for the DM's belief about the machine when deciding to act. In the first no-interaction setting, the DM always ignores the machine's prescription and bases the DM's choice solely on the DM's own judgment S_t^{H}. In this sense, the DM and the machine do not interact when deciding on tasks. In the second no-overriding

setting, the machine fully determines the DM's choice so that the DM never overrides its prediction S_t^{M}. Importantly, belief b_{t-1} does not determine whether an action is taken in both benchmarks and, thus, whether a machine's prediction is verified ex post. As a result, the second mechanism by which belief b_{t-1} influences posterior b_t in Equation (6) is mute. This belief affects learning through only the first mechanism, that is, the application of Bayes' rule when the DM acts.

More specifically, the DM acts if and only if $S_t^{\mathsf{H}} = +$ regardless of the machine's signal S_t^{M} in the no-interaction benchmark and if and only if $S_t^{\mathsf{M}} = +$ regardless of the DM's own judgment S_t^{H} in the no-overriding benchmark. The condition for acting, $\mathbb{P}(\Theta_t = \mathsf{A} \mid S_t^{\mathsf{H}}, S_t^{\mathsf{M}}, b_{t-1}) \geq r$, thus, reduces to $\mathbb{P}(\Theta_t = \mathsf{A} \mid S_t^{\mathsf{H}}) \geq r$ in the first benchmark and to $\mathbb{P}(\Theta_t = \mathsf{A} \mid S_t^{\mathsf{M}}, b_{t-1}) \geq r$ in the second one, which are, respectively, equivalent to $S_t^{\mathsf{H}} = +$ and to $S_t^{\mathsf{M}} = +$ for any b_{t-1} because of informativeness (1)–(3). In both benchmarks, Equation (6) then becomes

$$b_t = \begin{cases} b_{t-1} & \text{if } S_t^{\sigma} = - \\ \left[1 + \dfrac{\overline{b}_{t-1}}{b_{t-1}} \dfrac{\mathbb{P}^{\mathsf{W}}(S_t^{\mathsf{M}} \mid \Theta_t = \theta)}{\mathbb{P}^{\mathsf{B}}(S_t^{\mathsf{M}} \mid \Theta_t = \theta)}\right]^{-1} & \text{if } S_t^{\sigma} = +. \end{cases} \tag{7}$$

Here, $\sigma = \mathsf{H}$ and $\sigma = \mathsf{M}$ in the no-interaction and no-overriding benchmarks, respectively. In particular, the realization of S_t^{σ} is independent of belief b_{t-1}, which is in contrast to the condition in Equation (6).

To study the asymptotic behavior of b_t, we consider instead the log-likelihood ratio L_t of the probability that $\Gamma = \mathsf{B}$ in period t. Formally, L_t is a monotone continuous transformation of b_t given by $L_t \triangleq \log(b_t/(1-b_t))$, such that

$$L_t = L_{t-1} + R_t,$$

where $(R_t)_{t\in\mathbb{N}}$ are independent and identically distributed (i.i.d.) random jumps. In particular, the log-likelihood ratio is increasing in the DM's belief, and the asymptotic behavior of L_t fully determines the asymptotic behavior of b_t. Indeed, we have $b_t \xrightarrow{\text{a.s.}} 1$ (and $b_t \xrightarrow{\text{a.s.}} 0$) if and only if $L_t \xrightarrow{\text{a.s.}} +\infty$ (and $L_t \xrightarrow{\text{a.s.}} -\infty$) per the continuous mapping theorem.

Log-likelihood ratio L_t is a random walk governed by jumps $(R_t)_{t\in\mathbb{N}}$, which capture the magnitude and direction of the belief's updates. These random jumps take three possible values: a positive (negative) value when the DM observes that the machine's prediction is correct (wrong), that is, $S_t^{\mathsf{M}} = +$ and $\Theta_t = \mathsf{A}$ ($\Theta_t = \mathsf{NA}$), or zero when the task's type is not observed, that is, when $S_t^{\mathsf{H}} = -$ in the no-interaction benchmark and $S_t^{\mathsf{M}} = -$ in the no-overriding benchmark. The asymptotic behavior of L_t is then fully determined by the sign of the mean $\mathbb{E}^{\Gamma}[R_t]$. If $\mathbb{E}^{\Gamma}[R_t] > 0$ (<0), then log-likelihood ratio L_t increases in expectation and converges almost surely to $+\infty$ ($-\infty$),[6] whereas L_t does not converge when $\mathbb{E}^{\Gamma}[R_t] = 0$.

4.1. No-Interaction Benchmark

First, we characterize the DM's ability to learn the machine's type in the absence of DM–machine interactions, that is, when the DM's decisions always ignore the machine's prescriptions. Specifically, when Equation (7) holds with $\sigma = \mathsf{H}$, we have the following.

Theorem 1 (Learning with No Interaction). *When the machine is better* $\Gamma = \mathsf{B}$ *(is worse* $\Gamma = \mathsf{W}$*),* $b_t \xrightarrow{\text{a.s.}} 1$ *(and* $b_t \xrightarrow{\text{a.s.}} 0$*).*

To prove this result, we first establish that $\mathbb{E}^{\Gamma}[R_t] > 0$ (<0) if $\Gamma = \mathsf{B}$ ($\Gamma = \mathsf{W}$) and then apply the continuous mapping theorem. (All proofs are in the online appendix.)

Theorem 1 states that the DM correctly learns the machine's type when the DM decides to act solely based on the DM's own judgment. In particular, verification bias does not prevent learning in the limit. This is because the DM's sampling of the machine's correct and wrong predictions is not biased in this case. Indeed, the DM acts and, thus, verifies the machine when $\{S_t^{\mathsf{H}} = +\}$ regardless of the realization of S_t^{M}, and hence, the probability to verify is independent of the machine's prescription (conditional on the task's type). This, in effect, relaxes the exploration-free condition by randomly enabling learning (with probability $\mathbb{P}(S_t^{\mathsf{H}} = +)$) even when the machine's prescription would have induced the DM not to act (i.e., when $\mathbb{P}(\Theta_t = \mathsf{A} \mid S_t^{\mathsf{H}} = +, S_t^{\mathsf{M}} = -, b_{t-1}) < r$) in Equation (6).

Overall, the theorem reveals that inappropriate overriding decisions may occur only when the machine biases the DM's decisions because full learning occurs in the absence of DM–machine interactions. Next, we explore the case in which the machine fully biases these decisions and, hence, the sampling of observations.

4.2. No-Overriding Benchmark

In our next result, we characterize the DM's ability to learn the machine's type when the DM's choices are fully determined by the machine's prescriptions. In this case, the bias of the machine on the DM's decision is extreme. Specifically, we consider the DM's asymptotic learning behavior when Equation (7) holds with $\sigma = \mathsf{M}$. We then have the following.

Theorem 2 (Learning with No Overriding). *Unique thresholds* p^{B} *and* p^{W} *exist such that* $p^{\mathsf{B}} < p^{\mathsf{W}}$ *and*

- *When the machine is better* ($\Gamma = \mathsf{B}$), $b_t \xrightarrow{\text{a.s.}} 0$ *if* $p < p^{\mathsf{B}}$, $b_t \xrightarrow{\text{a.s.}} 1$ *if* $p > p^{\mathsf{B}}$ *and* b_t *is recurrent and oscillates if* $p = p^{\mathsf{B}}$.
- *When the machine is worse* ($\Gamma = \mathsf{W}$), $b_t \xrightarrow{\text{a.s.}} 0$ *if* $p < p^{\mathsf{W}}$, $b_t \xrightarrow{\text{a.s.}} 1$ *if* $p > p^{\mathsf{W}}$ *and* b_t *is recurrent and oscillates if* $p = p^{\mathsf{W}}$.

Further, we have $p^{\mathsf{B}} \triangleq \dfrac{\overline{\beta}^{\mathsf{B}} \log\left(\frac{\overline{\beta}^{\mathsf{W}}}{\overline{\beta}^{\mathsf{B}}}\right)}{\overline{\beta}^{\mathsf{B}} \log\left(\frac{\overline{\beta}^{\mathsf{W}}}{\overline{\beta}^{\mathsf{B}}}\right) + \alpha^{\mathsf{B}} \log\left(\frac{\alpha^{\mathsf{B}}}{\alpha^{\mathsf{W}}}\right)}$ *and* $p^{\mathsf{W}} \triangleq \dfrac{\overline{\beta}^{\mathsf{W}} \log\left(\frac{\overline{\beta}^{\mathsf{W}}}{\overline{\beta}^{\mathsf{B}}}\right)}{\overline{\beta}^{\mathsf{W}} \log\left(\frac{\overline{\beta}^{\mathsf{W}}}{\overline{\beta}^{\mathsf{B}}}\right) + \alpha^{\mathsf{W}} \log\left(\frac{\alpha^{\mathsf{B}}}{\alpha^{\mathsf{W}}}\right)}$.

In essence, Theorem 2 states that the DM's ability to learn depends on the DM's prior about the task as well as the machine's type. The DM learns that the machine is worse (better) when the DM's prior is below (above) p^Γ for type $\Gamma \in \{\mathsf{B}, \mathsf{W}\}$. Importantly, this means that the DM may not properly learn whether the machine is better than the DM. Indeed, when prior p is low ($p < p^{\mathsf{B}}$), the DM learns that the machine is worse ($b_t \xrightarrow{\text{a.s.}} 0$), whereas the machine is actually better ($\Gamma = \mathsf{B}$). Similarly, when prior p is high ($p > p^{\mathsf{W}}$), the DM learns that the machine is better ($b_t \xrightarrow{\text{a.s.}} 1$), whereas the machine is actually worse ($\Gamma = \mathsf{W}$).

Theorem 2 stems from the fact that, in this benchmark, the DM acts and observes the machine's correctness only if the machine's signal is positive. In other words, the DM's observations are sampled solely from true and false positive predictions and never from true or false negative ones. Indeed, recall first that the DM's belief increases (decreases) when the DM observes a correct (incorrect) machine recommendation. Because the DM is able to observe this only when the machine's signal is positive, the DM increases the DM's belief if and only if the machine correctly prescribes to act ($S_t^{\mathsf{M}} = +$ and $\Theta_t = \mathsf{A}$) and decreases the DM's belief if and only if the machine wrongly prescribes to act ($S_t^{\mathsf{M}} = +$ and $\Theta_t = \mathsf{NA}$). But whether the task truly requires an action (i.e., $\Theta_t = \mathsf{A}$) is determined by prior p. Hence, the DM increases the DM's belief more frequently when the task is more likely to require an action, that is, prior p takes higher values. The DM's belief converges to one (to zero) when prior p is sufficiently high (low) such that the number of observed correct predictions is relatively higher than the number of incorrect ones.

This effect of prior p is absent from the no-interaction benchmark because the DM also observes the correctness of the machine's predictions when the machine recommends not to act (i.e., when $S_t^{\mathsf{M}} = -$ and $\Theta_t = \mathsf{A}$ or $S_t^{\mathsf{M}} = -$ and $\Theta_t = \mathsf{NA}$).

Note, finally, that magnitudes of these changes in beliefs do not depend on prior p but are determined by the accuracy parameters of the machine. Threshold p^Γ, thus, corresponds to the break-even value of prior p such that the expected increase in belief compensates for the expected decrease when the machine is of type Γ. When $p > p^\Gamma$, the expected increase dominates the expected decrease, and the belief converges to one. When $p < p^\Gamma$, the opposite is true, and the belief converges to zero.

5. Main Setup: Accounting for the DM's Belief About the Machine

In our main setup, and in contrast to the previous benchmarks, the DM's belief about the machine influences the DM's decision to act and, hence, the DM's ability to verify the machine's predictions. As a result, learning is endogenously determined by the DM's current belief about the machine. This fundamentally changes the nature of mislearning.

More specifically, recall that, because of informativeness, the DM always decides according to the DM's signal when it is consistent with the machine's signal with $S_t^{\mathsf{H}} = S_t^{\mathsf{M}}$. When these signals differ with $S_t^{\mathsf{H}} \neq S_t^{\mathsf{M}}$, the DM may override the machine when the DM's current belief b_{t-1} that the machine is better is sufficiently low. Hence, belief b_{t-1} influences posterior b_t via the two mechanisms captured by Equation (6). The following result determines when such overriding decisions occur. (The result follows from Substitutions (4) and (5) and the continuity of the posterior probabilities in b_{t-1}; see Online Appendix B.)

Lemma 1. *Unique thresholds* $b^- \in (0,1)$ *and* $b^+ \in (0,1)$ *exist such that*

$$\mathbb{P}(\Theta_t = \mathsf{A} \mid S_t^{\mathsf{H}} = +, S_t^{\mathsf{M}} = -, b_{t-1}) \geq r \Leftrightarrow b_{t-1} \leq b^-, \tag{8}$$

$$\mathbb{P}(\Theta_t = \mathsf{A} \mid S_t^{\mathsf{H}} = -, S_t^{\mathsf{M}} = +, b_{t-1}) \leq r \Leftrightarrow b_{t-1} \leq b^+. \tag{9}$$

Lemma 1 states that, when the DM's judgment contradicts the machine's prescription, that is, $S_t^{\mathsf{H}} \neq S_t^{\mathsf{M}}$, the DM overrides the machine if and only if the DM's belief in a better machine is sufficiently low, that is, below a threshold. However, the DM can override the machine in two different ways, depending on whether the machine prescribes to act or not. This yields two different thresholds b^- and b^+, which correspond to an overriding decision for a negative and positive machine signal, respectively.

These thresholds actually correspond to the value of belief b that makes the DM indifferent between acting and not acting when $S_t^{\mathsf{H}} = -, S_t^{\mathsf{M}} = +$ and $S_t^{\mathsf{H}} = +, S_t^{\mathsf{M}} = -$, respectively. Note also that the ranking between b^- and b^+ depends on the problem's parameters, and we define the minimum and maximum of these two thresholds as $b^{\mathsf{H}} \triangleq \min(b^-, b^+)$ and $b^{\mathsf{M}} \triangleq \max(b^-, b^+)$, respectively (where b^{H} and b^{M} can be equal).

Thus, when belief b_{t-1} is sufficiently large with $b_{t-1} > b^{\mathsf{M}}$, the DM has sufficient confidence in the machine to always follow its prescriptions; this corresponds to the no-overriding benchmark. However, when the belief is sufficiently low with $b_{t-1} < b^{\mathsf{H}}$, the DM always overrides the machine and decides solely based on the DM's judgment, which corresponds to the no-interaction benchmark. Overall, these two cases are consistent with substitution, which stipulates that the machine is either better or worse than the DM.

Interestingly, Lemma 1 further reveals that the DM may treat the machine's prescription as complementing—instead of substituting—the DM's expertise. This occurs when the DM is sufficiently unsure about the machine's type with $b_{t-1} \in (b^{\mathsf{H}}, b^{\mathsf{M}})$. In this case, the DM and the machine complement one another in two possible ways, depending on whether threshold b^- is larger or smaller than threshold

b^+. If $b^+ < b^-$, the DM overrules the machine when its signal is negative but follows the machine's prescription when it is positive. In other words, the DM assumes that the DM makes better positive but worse negative decisions than the machine. In this sense, the DM and the machine collaborate on the task. As a result, the DM acts if and only if either the DM or the machine find evidence to do so ($S_t^{\mathsf{H}} = +$ or $S_t^{\mathsf{M}} = +$). If $b^- < b^+$, however, the DM overrules a positive machine's signal but follows a negative machine's signal and, thus, acts if and only if the DM and the machine agree that an action is required ($S_t^{\mathsf{H}} = +$ and $S_t^{\mathsf{M}} = +$).

Overall, Lemma 1 indicates that the DM's ability to learn the true type of task depends on the DM's current belief about the machine. This means, in particular, that the random jumps of the corresponding log-likelihood ratio also depend on the current ratio. Formally, we have

$$L_t = L_{t-1} + R_t^{\mathsf{HM}}(L_{t-1})$$

when the DM can override the machine. In contrast to the no-overriding benchmark, the random jumps R_t^{HM} are no longer i.i.d. as their distributions now depend on the magnitude of L_{t-1}. Thus, the sign of the expected jump, which determines the asymptotic behavior of belief b_t, is path-dependent. Next, we explore how this dependency affects the ability of the DM to learn the true machine type.

5.1. Learning When the Machine Is Better

We first study the DM's ability to properly learn the machine's type when the machine is, in fact, better. Our next result characterizes the situations in which mislearning occurs in this case.

Theorem 3. *When the machine is better, that is, $\Gamma = \mathsf{B}$, if $p \le p^B$, then b_t oscillates and is recurrent; otherwise, $b_t \xrightarrow{a.s.} 1$.*

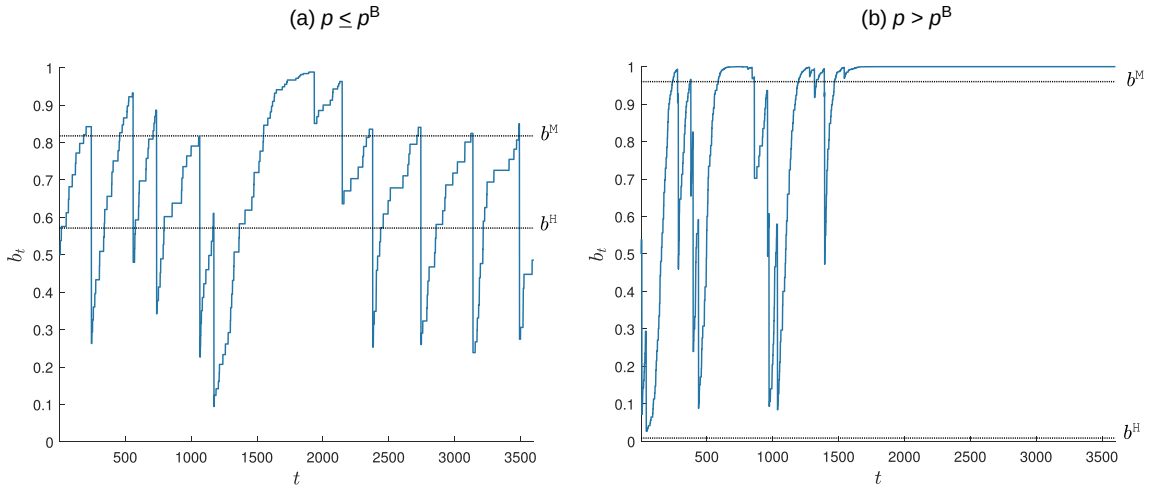
Thus, the DM's ability to learn the machine's type continues to depend on whether the DM's prior about the task is sufficiently high. In fact, the threshold characterizing when proper learning occurs is the exact same as the one without overriding (defined in Theorem 2). Specifically, the DM learns that the machine is indeed better ($b_t \xrightarrow{a.s.} 1$) if and only if prior p is sufficiently high with $p > p^{\mathsf{B}}$. Figure 1(b) illustrates this case and exhibits asymptotic behavior.

The DM's ability to override the machine, however, fundamentally changes the nature of mislearning. Indeed, when prior p is such that $p \le p^{\mathsf{B}}$, the DM wrongly learns that the machine is worse in the no-overriding benchmark. In the main setup, however, the belief oscillates as illustrated in Figure 1(a). Additionally, because the belief is also recurrent, the DM constantly switches among overruling the machine ($b_t < b^{\mathsf{H}}$), collaborating with it ($b_t \in (b^{\mathsf{H}}, b^{\mathsf{M}})$), or letting the machine decide ($b_t > b^{\mathsf{M}}$) as stated by the following corollary.

Corollary 1. *When the machine is better, that is, $\Gamma = \mathsf{B}$ and $p \le p^B$, intervals $(0, b^H]$, (b^H, b^M) and $[b^M, 1)$ are recurrent for belief b_t.*

Hence, when the DM sufficiently believes that the machine is better, the DM never overrules it, and we retrieve the dynamics of the no-overriding benchmark. That is, when $b_t > b^{\mathsf{M}}$, learning is entirely driven by whether a machine's prescription to act is correct. Additionally, because prior p is low, the frequency of these correct predictions is also low, so the belief is decreasing in expectation.

Figure 1. (Color online) The DM's Belief b_t When the Machine Is Better, $\Gamma = \mathsf{B}$



Note. $\alpha^{\mathsf{H}} = \beta^{\mathsf{H}} = 0.95$, $\alpha^{\mathsf{B}} = \beta^{\mathsf{B}} = 0.99$, $\alpha^{\mathsf{W}} = \beta^{\mathsf{W}} = 0.85$, $p^{\mathsf{B}} = 0.15$ and $r = 0.07$; (a) $p = 0.05$, $b^{\mathsf{H}} = 0.57$, $b^{\mathsf{M}} = 0.81$; (b) $p = 0.2$, $b^{\mathsf{H}} = 0.01$, $b^{\mathsf{M}} = 0.96$.

In contrast, when the DM sufficiently believes that the machine is worse with $b_t < b^{\mathrm{H}}$, the DM always overrides the machine. In this case, the DM sometimes observes the machine's accuracy even when it prescribes not to act. This occurs when the DM's signal is positive and overrules a machine's negative signal. In this case, learning is driven by the true machine type, and because the machine is actually better, the belief increases in expectation.

Overall, the result holds because the DM's belief in the machine's type negatively reinforces the DM's sampling bias: when the belief biases the sample of observations, the resulting updated belief tends to debias the sampling and vice versa. As a result, belief b_t is pushed back downward when it reaches high values ($b_t > b^{\mathrm{M}}$) and is pushed upward when it takes low values ($b_t < b^{\mathrm{H}}$). Hence, the DM never fully learns that the machine is better but, because of overriding, never mislearns that it is worse either. In this sense, the DM always remains in perpetual uncertainty about whether to disregard the machine.

Interestingly, this long-run uncertainty induces the DM to sometimes treat the machine's prescription as a complement to the DM's judgment. This happens when the belief reaches $b_t \in (b^{\mathrm{H}}, b^{\mathrm{M}})$, which is a recurrent event. In this case, the DM and the machine coproduce the decision per Lemma 1 (and the explanations that follow). Because the machine and DM are actually substitutes, the emergence of this complementarity is driven only by the DM's inability to learn the true machine type.

5.2. Learning When the Machine Is Worse

Per Theorems 2 and 3, the DM properly learns that the machine is good when prior p takes high values (i.e., $p > p^{\mathrm{B}}$), whether the DM can override the machine or not. In this case, overriding essentially prevents the DM from wrongly learning that the machine is worse, which creates a perpetual state of uncertainty. In contrast, when the machine is indeed worse and the DM can overrule it, the DM may learn its true type for any prior p. This, however, occurs only randomly when prior p takes low values as stated by the following result.

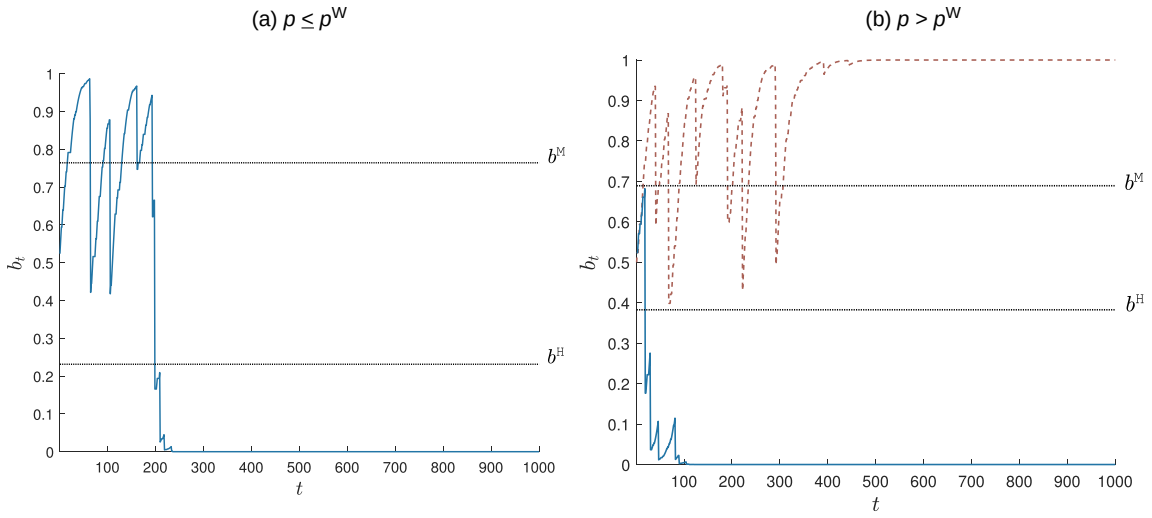
Theorem 4. *When the machine is worse, that is, $\Gamma = W$, if $p \le p^{W}$, then $b_t \xrightarrow{a.s.} 0$; otherwise, $b_t \xrightarrow{a.s.} X$, where X is a Bernoulli random variable.*

Theorem 4 indicates that the DM's ability to learn hinges again on prior p. As in the no-overriding benchmark, the DM can properly learn that the machine is worse ($b \xrightarrow{a.s.} 0$) if prior p takes low values ($p < p^{\mathrm{W}}$, where threshold p^{W} is, again, the same as that in the no-overriding benchmark). Figure 2(a) illustrates this point and depicts a sample path of b_t.

When the prior is high ($p > p^{\mathrm{W}}$), however, the belief converges to a Bernoulli random variable. That is, the sample paths of belief b_t converge to zero with a certain probability and to one with the complement probability. In particular, the belief never oscillates or converges to an interior point in the limit. Thus, the DM's ability to properly learn the machine's type is random in this case. In particular, the DM may sometimes wrongly learn that the machine is better, whereas it is actually worse. Figure 2(b) illustrates this point and depicts examples of the two possible sample paths for b_t, one (dashed line) converging to one and the other (solid line) to zero.

Similar to the better machine case, learning is driven by prior p as in the no-overriding benchmark when the belief

Figure 2. (Color online) The DM's Belief b_t When the Machine Is Worse, $\Gamma = \mathrm{W}$



Note. $\alpha^{\mathrm{H}} = \beta^{\mathrm{H}} = 0.95$, $\alpha^{\mathrm{B}} = \beta^{\mathrm{B}} = 0.99$, $\alpha^{\mathrm{W}} = \beta^{\mathrm{W}} = 0.9$, $p^{\mathrm{W}} = 0.72$ and $r = 0.8$; (a) $p = 0.71$, $b^{\mathrm{H}} = 0.23$, $b^{\mathrm{M}} = 0.76$; (b) $p = 0.75$, $b^{\mathrm{H}} = 0.38$, $b^{\mathrm{M}} = 0.68$.

is high ($b_t > b^M$) and by the true type of the machine when the belief is low ($b_t < b^H$). In the latter case, the belief decreases in expectation because the machine is worse.

Thus, for low prior $p < p^W$, the belief moves downward in expectation when it takes sufficiently high or low values and, hence, converges to zero in the long run. The DM then properly learns that the machine is worse.

For high prior $p > p^W$, however, the belief increases in expectation when the belief is already high and decreases when it is already low. In the long run, the belief is, thus, pushed close to either one or zero. This is because, in contrast to the case in which the machine is better, the DM's belief positively reinforces the sampling bias: When the belief biases (does not bias) the sample, the resulting updated belief tends to perpetuate the bias (remain unbiased). Whether the belief reaches high or low values is then determined by realizations of the different signals and the task types and is, thus, random. Note that when the belief takes intermediary values ($b_t \in (b^H, b^M)$) it can either decrease or increase in expectation depending on the problem parameters. However, this region is transient because the belief is pushed away from the region when the belief is more extreme ($b_t \notin (b^H, b^M)$).

6. Implications

6.1. Learning and Mislearning

Taken together, these results provide theoretical limits on our ability to learn whether a machine makes better decisions than an expert. Interestingly, this inability to learn sometimes induces the DM to treat the machine's prescription as a complement to the DM's own judgment even though the two are actually substitutes. For instance, the DM may believe that the DM's predictions have better sensitivity but worse specificity than those of the machine, whereas in fact, the machine is better in terms of both accuracy metrics. In this sense, the DM's uncertainty about the machine provides a novel rationale for why experts and machines may collaborate in practice.

Our results further identify the uncertainty surrounding the decision task as the key factor for mislearning. In fact, the DM fails to learn when the DM is most certain a priori about whether an action is required for a task (i.e., when prior p takes more extreme values with $p \le p^B$ or $p > p^W$). Conversely, the DM always properly learns the machine's type when the DM is most uncertain about whether to act (i.e., prior p takes moderate values) as stated by the next corollary of Theorems 3 and 4.

Corollary 2. *The DM always correctly learns the type of the machine if and only if* $p \in (p^B, p^W]$.

6.2. Learning with Anticipation

In our setup, as in the literature, the DM updates the DM's belief using the past history of the observed accuracy of the machine's predictions. Nonetheless, our results characterize the asymptotic behavior of this learning process and, as such, provide guidelines for DMs who anticipate the future behavior of their own belief. In particular, the nature of a learning failure is indicative of the machine's type in our results. The DM may, thus, leverage this information to determine whether the machine is better.

Indeed, the DM's belief may oscillate only when the machine is better (see Figure 1(a)) and always converges when it is worse (see Figure 2). Thus, the longer the DM remains uncertain (in the sense of Theorem 3), the more likely it is that the machine is actually better. Similarly, the DM's belief can converge to zero only if the machine is worse. Indeed, the belief either oscillates or converges to one when the machine is better. Hence, the longer the DM believes that the machine is worse, the more likely the DM is correct in the DM's assessment.

Assessing if the DM is correct when the DM increasingly believes that the machine is better appears to be more challenging. Indeed, the DM's belief can converge to one whether the machine is better (see Figure 1(b)) or worse (see Figure 2(b)). To circumvent this issue, one approach consists of relying on more than one decision maker. To see how, consider several identical decision makers who independently handle a series of tasks that are randomly drawn from the same sample and use the same machine. If this machine is better, all DMs should have the same limiting behavior: they either all remain uncertain or all learn that the machine is indeed better (per Theorem 3). However, if the machine is worse, the convergence to either zero or one is random (per Theorem 4). Thus, if a single DM in the team believes over time that the machine is worse, then the machine is indeed likely to be worse even if the rest of the team believes it to be better. In contrast, if there is consensus in the team that the machine is better, then the larger the team is, the more likely it is that the machine is better.

In short, long-term uncertainty or a unanimous belief among large teams that the machine is better is indicative of a better machine. In contrast, persistently overruling the machine is indicative of a worse one.

6.3. Adoption or Rejection

Our study also sheds lights on the decision to fully adopt or reject the machine. Indeed, after observing and, at times, overriding the machine's prescriptions, the DM's belief may reach extreme levels. In these cases, the DM decides either to let the machine make all the decisions (as in Section 4.2) or to abandon the machine altogether, depending on whether the belief is sufficiently high or low, respectively. Once a machine is abandoned, however, the DM cannot learn about it anymore.

If the DM decides to fully adopt the machine—but continues to observe its performance—our results indicate

that the DM becomes increasingly confident about the DM's adoption decision over time when prior p about the task is high ($p > p^{\mathsf{B}}$ for a better machine and $p > p^{\mathsf{W}}$ for a worse one per Theorems 3 and 4). This occurs even when the machine is actually worse and should be abandoned.

In contrast, when the prior about the task is low ($p < p^{\mathsf{B}}$ or $p < p^{\mathsf{W}}$, depending on the true machine type), the DM increasingly doubts the DM's adoption decision over time. This is because the DM's belief in a better machine decreases in expectation over time and always approaches zero in the limit in this case (per Theorems 3 and 4). This happens even when the machine is actually better and should be adopted.

Recall, finally, that when the machine is better and the prior about the task is low, the DM's belief in the main setup oscillates and is recurrent in (0,1) (per Theorem 3 and Corollary 1). Therefore, the DM's belief eventually reaches any low level with probability one. In other words, the DM always ends up abandoning the machine in the long run even though the machine is actually better.

7. Mistrust Biases Against the Machine

A key feature of the mislearning behavior in Theorems 3 and 4 is that they do not stem from an inherent mistrust against the machine. Instead, they stem from four fundamentals (verification bias, exploration-free decisions, informativeness, and substitution), which characterize the setup in which the DM works with the machine. Nonetheless, the DM may also be subject to mistrust biases against the machine in situations in which these fundamentals are at play. In this section, we explore to what extent biases such as these interact with our four fundamentals to affect the DM's learning behavior.

In our main setup, mistrusting the machine affects the DM's ability to learn in at least two ways. First, the DM may downplay the machine's prescription when deciding to act, which alters the DM's ability to observe the correctness of the machine's predictions. Second, and in line with the algorithm aversion reported by Dietvorst et al. (2015), the DM's belief in the machine may disproportionately drop upon observing the failure of a machine's prediction. In the following, we inspect these different biases in turn.

7.1. Mistrusting the Machine When Deciding

The DM's mistrust in the machine may affect the way the DM weights the machine's prescription when deciding to act. This is consistent with the decision-making literature, which finds that individuals tend to discount information coming from external sources and overweight their own opinions (see, for instance, Soll and Mannes 2011). To account for this possibility, we follow Stone (1961), who proposes a non-Bayesian approach to represent the aggregation of different opinions. This approach is commonly used to model mistrust bias (Özer and Zheng 2018), in particular, when a human makes decisions based on the input of a data-driven algorithm (Ahsen et al. 2019, Boyacı et al. 2023). Specifically, the DM's updated belief that an action is required given the DM's and machine's signals is a linear combination between the updated belief of the human and that of the machine. More formally, the DM's posterior belief about the task is defined as

$$\begin{aligned}&\tilde{\mathbb{P}}_\lambda(s^{\mathsf{H}}, s^{\mathsf{M}}, b_{t-1})\\ &\triangleq \lambda\mathbb{P}(\Theta_t = \mathsf{A} \mid S^{\mathsf{H}} = s^{\mathsf{H}}) + (1-\lambda)\mathbb{P}(\Theta_t = \mathsf{A} \mid S^{\mathsf{M}} = s^{\mathsf{M}}, b_{t-1}),\end{aligned} \tag{10}$$

where $\lambda \in (0,1)$ represents the DM's mistrust bias against the machine's signal. The higher the value of λ, the more the DM mistrusts the machine. Belief $\tilde{\mathbb{P}}_\lambda(S^{\mathsf{H}}, S^{\mathsf{M}}, b_{t-1})$ corresponds then to the posterior probability $\mathbb{P}(\Theta = \mathsf{A} \mid S^{\mathsf{H}}, S^{\mathsf{M}}, b_{t-1})$ derived from Bayes' rule in the main setup. In particular, the DM decides to act if and only if $\tilde{\mathbb{P}}_\lambda(s^{\mathsf{H}}, s^{\mathsf{M}}, b_{t-1}) \geq r$.

In this setup, the DM always overrules a better machine (never overrules a worse machine) if the bias is too high (too low). In these extreme situations, the DM and the machine no longer substitute one another. We, thus, restrict our analysis to moderate values of mistrust parameter λ as formalized by the following lemma in which $\tilde{\mathbb{P}}^{\mathsf{B}}_\lambda(s^{\mathsf{H}}, s^{\mathsf{M}}) \triangleq \tilde{\mathbb{P}}_\lambda(s^{\mathsf{H}}, s^{\mathsf{M}}, 1)$ and $\tilde{\mathbb{P}}^{\mathsf{W}}_\lambda(s^{\mathsf{H}}, s^{\mathsf{M}}) \triangleq \tilde{\mathbb{P}}_\lambda(s^{\mathsf{H}}, s^{\mathsf{M}}, 0)$ represent the DM's beliefs about the task's type when the machine's type is known.

Lemma 2. *Two thresholds $\lambda_{\min}$ and $\lambda_{\max}$ exist such that, if $\lambda \in (\lambda_{\min}, \lambda_{\max})$, then*

$$\tilde{\mathbb{P}}^{B}_\lambda(S^H_t = +, S^M_t = -) < r \text{ and } \tilde{\mathbb{P}}^{B}_\lambda(S^H_t = -, S^M_t = +) \geq r, \tag{11}$$

$$\tilde{\mathbb{P}}^{W}_\lambda(S^H_t = +, S^M_t = -) \geq r \text{ and } \tilde{\mathbb{P}}^{W}_\lambda(S^H_t = -, S^M_t = +) < r. \tag{12}$$

In the following, we focus on $\lambda \in (\lambda_{\min}, \lambda_{\max})$ to ensure that substitution persists in the presence of mistrust bias. The next theorem shows that, under these conditions, the structure of our main results continue to hold.

Theorem 5. *Assume $\lambda \in (\lambda_{\min}, \lambda_{\max})$.*

- *When $\Gamma = B$, if $p \leq p^B$, then b_t oscillates and is recurrent; otherwise, $b_t \xrightarrow{a.s.} 1$.*
- *When $\Gamma = W$, if $p \leq p^W$, then $b_t \xrightarrow{a.s.} 0$; otherwise, $b_t \xrightarrow{a.s.} X$, where X is a Bernoulli random variable.*

Thus, the DM's learning behavior characterized in Theorems 3 and 4 does not change overall if the DM is biased against the machine's prescription when making a decision. Note also that Corollary 1 continues to hold in this case but with thresholds b^- and b^+ depending on λ (see Lemma 5 in the online appendix).

7.2. Mistrusting the Machine When Updating Belief

The DM's mistrust against the machine can also affect the way the DM updates the DM's belief about the machine. Dietvorst et al. (2015), for instance, experimentally show that individuals are more likely to ignore algorithm-based predictions after observing these algorithms err. More generally, the observation of negative outcomes, such as a prediction failure, more strongly impact the formation of an individual impression than do positive ones—a phenomenon referred to as negativity bias in the literature (Baumeister et al. 2001).

To account for this bias, we follow the literature (see, for instance, Coutts 2019, Möbius et al. 2022) and allow updated belief b_t to drop significantly upon observing an incorrect machine prediction. Specifically, the DM updates the DM's belief following Bayes' rule when the machine is correct but magnifies the decrease in belief when the machine is wrong. More formally, we have

$$b_t = \begin{cases} b_{t-1} & \text{if } \mathbb{P}(\Theta_t = \mathsf{A} \mid S_t^{\mathsf{H}} = s^{\mathsf{H}},\ S_t^{\mathsf{M}} = s^{\mathsf{M}},\ b_{t-1}) < r \\ \left[1 + \dfrac{\overline{b}_{t-1}}{b_{t-1}} \left[\dfrac{\mathbb{P}^{\mathsf{W}}(S_t^{\mathsf{M}} = s^{\mathsf{M}} \mid \Theta_t = \theta)}{\mathbb{P}^{\mathsf{B}}(S_t^{\mathsf{M}} = s^{\mathsf{M}} \mid \Theta_t = \theta)}\right]^{\phi(s^{\mathsf{M}},\theta)}\right]^{-1} & \\ \qquad \text{if } \mathbb{P}(\Theta_t = \mathsf{A} \mid S_t^{\mathsf{H}} = s^{\mathsf{H}},\ S_t^{\mathsf{M}} = s^{\mathsf{M}},\ b_{t-1}) \geq r, \end{cases}$$

where function $\phi(s^{\mathsf{M}}, \theta) = \mu > 1$ if the machine is incorrect (i.e., for $s^{\mathsf{M}} = +$ and $\theta = \mathsf{NA}$ or $s^{\mathsf{M}} = -$ and $\theta = \mathsf{A}$) and $\phi(s^{\mathsf{M}}, \theta) = 1$ otherwise. Because ratio $\mathbb{P}^{\mathsf{W}}(S_t^{\mathsf{M}} | \Theta_t)/\mathbb{P}^{\mathsf{B}}(S_t^{\mathsf{M}} | \Theta_t) > 1$ when the machine is incorrect, the higher the value of mistrust parameter μ is, the lower belief b_t becomes. In particular, the main setup corresponds to $\mu = 1$, which coincides with Bayes' rule.

The next result characterizes the asymptotic behavior of the DM's belief in the presence of this negativity bias.

Theorem 6 (Learning with Negativity Bias). *Unique thresholds μ^B, μ^W, μ^H exist such that*

- *When the machine is better ($\Gamma = B$),*
  - *If $\mu \geq \mu^B$ and $\mu > \mu^H$, then $b_t \xrightarrow{a.s.} 0$.*
  - *If $\mu^B > \mu > \mu^H$, then $b_t \xrightarrow{a.s.} X$, where X is a Bernoulli random variable.*
  - *If $\mu^H \geq \mu \geq \mu^B$, then b_t is recurrent and oscillates.*
  - *If $\mu^B > \mu$ and $\mu^H \geq \mu$, then $b_t \xrightarrow{a.s.} 1$.*
- *When the machine is worse ($\Gamma = W$),*
  - *If $\mu \geq \mu^W$, then $b_t \xrightarrow{a.s.} 0$.*
  - *If $\mu^W > \mu$, then $b_t \xrightarrow{a.s.} X$, where X is a Bernoulli random variable.*

Further, we have

$$\mu^B \triangleq \frac{p\alpha^B \log\left(\frac{\alpha^B}{\alpha^W}\right)}{\overline{p}\overline{\beta}^B \log\left(\frac{\overline{\beta}^W}{\overline{\beta}^B}\right)},\ \mu^W \triangleq \frac{p\alpha^W \log\left(\frac{\alpha^B}{\alpha^W}\right)}{\overline{p}\overline{\beta}^W \log\left(\frac{\overline{\beta}^W}{\overline{\beta}^B}\right)} \quad \textit{and}$$

$$\mu^H \triangleq \frac{p\alpha^H\alpha^B \log\left(\frac{\alpha^B}{\alpha^W}\right) + \overline{p}\overline{\beta}^H\beta^B \log\left(\frac{\beta^B}{\beta^W}\right)}{p\alpha^H\overline{\alpha}^B \log\left(\frac{\overline{\alpha}^W}{\overline{\alpha}^B}\right) + \overline{p}\overline{\beta}^H\overline{\beta}^B \log\left(\frac{\overline{\beta}^W}{\overline{\beta}^B}\right)} > 1.$$

Note that thresholds μ^{B} and μ^{W} actually play the same role as p^{B} and p^{W} in Theorems 3 and 4, respectively. Indeed, thresholds p_μ^{B} and p_μ^{W} exist such that $\mu > \mu^\Gamma \Leftrightarrow p < p_\mu^\Gamma$ for $\Gamma = \{\mathsf{B}, \mathsf{W}\}$. In particular, the structure of the results when the machine is worse (see Theorem 4) does not change in the presence of negativity bias. In this case, mistrust parameter μ affects the learning only through the value of threshold μ^{W} and, hence, p_μ^{W}.

When the machine is better, however, the presence of mistrust changes the structure of the results. In particular, under moderate mistrust such that $\mu^{\mathsf{B}} > \mu > \mu^{\mathsf{H}}$, the DM learns that the machine is better only with some probability (second point in Theorem 4). This is in contrast to the main setup without mistrust, in which the DM always learns that the machine is better if $p > p^{\mathsf{B}}$. In fact, the DM's belief can converge to a Bernoulli random variable in our main setup only when the machine is worse. If the mistrust in the machine is too strong with $\mu > \max(\mu^{\mathsf{H}}, \mu^{\mathsf{B}})$ (first point of the theorem), however, the DM always wrongly learns that the machine is worse. Otherwise, the bias does not alter the learning behavior. In fact, Theorem 6 reduces to Theorems 3 and 4 when $\mu = 1$. In this case, we have $\mu^{\mathsf{H}} > \mu = 1$, and the belief either converges to one or oscillates depending on whether $\mu^{\mathsf{B}} \leq \mu = 1$ or not, which is equivalent to $p^{\mathsf{B}} \geq p$.

Overall, mistrust in the form of a negativity bias interacts with our fundamentals in a meaningful way only when the level of mistrust is moderate and the machine is actually better. In this case, whether the DM learns the true nature of the machine becomes random, whereas the DM always properly learns that the machine is better when the DM is not biased against the machine.

8. Complementarity

Thus far, we focus on settings in which the machine and the DM are substitutes. Nonetheless, our framework also applies to the case of complementarity, which can take different forms. In this section, we explore the learning behavior of a DM who uncovers how a machine may complement the DM's own judgment.

In our setup, only two possible ways actually exist by which the machine and the DM complement one another. Indeed, a machine that complements the DM

is superior in only one of the two dimensions of a judgment—positive or negative signals—and inferior in the other. Thus, the first form of complementarity corresponds to a DM who always overrides the machine when the DM's judgment indicates that an action is required but always follows the machine's prescription if the DM finds that the DM should not act. The converse holds for the second form: the DM overrides the machine when the DM's judgment indicates not to act but always follows the machine otherwise.

We denote these two machine types as $\mathsf{C}+$ and $\mathsf{C}-$, respectively, and their sensitivity and specificity are α^Γ and β^Γ for $\Gamma \in \{\mathsf{C}+, \mathsf{C}-\}$. In this section, we study the DM's learning behavior in the same settings as our base model except for the Substitutions (4) and (5), which we replace by the following complementarity conditions.

8.1. Complementarity

$$\mathbb{P}^{\mathsf{C}+}(\Theta = \mathsf{A} \mid S^{\mathsf{H}} = +, S^{\mathsf{M}} = -) < r \text{ and}$$
$$\mathbb{P}^{\mathsf{C}+}(\Theta = \mathsf{A} \mid S^{\mathsf{H}} = -, S^{\mathsf{M}} = +) < r, \tag{13}$$
$$\mathbb{P}^{\mathsf{C}-}(\Theta = \mathsf{A} \mid S^{\mathsf{H}} = +, S^{\mathsf{M}} = -) \geq r \text{ and}$$
$$\mathbb{P}^{\mathsf{C}-}(\Theta = \mathsf{A} \mid S^{\mathsf{H}} = -, S^{\mathsf{M}} = +) \geq r, \tag{14}$$

where $\mathbb{P}^{\mathsf{C}+}\{\cdot\}$ and $\mathbb{P}^{\mathsf{C}-}\{\cdot\}$ denote the probability measures induced by the two types of the machine.

The DM does not know the machine's type a priori. However, the DM forms a belief over time about which type of complementarity the DM is facing. With a slight abuse of notation, we refer to b_t as the DM's prior belief that $\Gamma = \mathsf{C}+$. The next result then characterizes the DM's ability to learn how the machine complements the DM's judgment.

Theorem 7. *We have*

- *When* $\Gamma = \mathsf{C}+$, *then* $b_t \xrightarrow{a.s.} 1$.
- *When* $\Gamma = \mathsf{C}-$, *a unique threshold* p^{C} *exists such that* $b_t \xrightarrow{a.s.} 1$ *if* $p \leq p^{\mathsf{C}}$; *otherwise,* $b_t \xrightarrow{a.s.} X$, *where* X *is a Bernoulli random variable. (Threshold* p^{C} *is defined in* (80) *in Online Appendix D.)*

Thus, the DM always properly learns the machine's type when the actual form of complementarity is $\mathsf{C}+$. In contrast, the DM can mislearn how the machine complements the DM's judgment when the true type is $\mathsf{C}-$ and the prior about the task is high (i.e., $p > p^{\mathsf{C}}$). In this case, learning is random, and the DM wrongly learns with positive probability that the machine is of type $\mathsf{C}+$.

This result is akin to Theorem 4 when the machine and DM are substitutes. In contrast to Theorem 3, however, the DM's belief never oscillates and always converges to either zero or one. Thus, in the limit, the DM is always certain of the form of complementarity that the machine provides. In particular, the DM never behaves as if the machine and DM were substitutes. Again, this is in contrast to our main setup, in which the DM's decisions sometimes exhibit complementarity, whereas in fact, the DM and the machine are substitutes (see Corollary 1).

Note finally that Conditions (13) and (14) correspond to a complementarity between the machine's and the DM's signals. Other forms of complementarity, however, exist. In particular, the DM may seek to uncover for which decision tasks the machine is better and for which ones the DM is. In the context of biopsies, for instance, this corresponds to understanding for what kinds of patients the machine does better and for what kinds of patients it does worse. A possible approach to study this case is to consider our main setup but with more than one type of decision task. Denote this type as T with threshold r^{T} and prior p^{T} for $\mathsf{T} \in \{\mathsf{T}_1, \mathsf{T}_2, \dots\}$. These different types may correspond to different kinds of patients, for instance. The DM then forms different beliefs b_t^{T} over time so that the findings of Section 5 independently apply to each task's type $\mathsf{T} \in \{\mathsf{T}_1, \mathsf{T}_2, \dots\}$. With multiple task types, these findings characterize when the DM wrongly learns the decision tasks for which the machine's predictions are superior, and the ones for which the DM's own judgment is better.

9. Partial Relaxation of the Verification Bias

In this section, we explore the robustness of our results when the verification bias is partially relaxed, which is legitimate when the bias stems from the DM's limited attention. In this context, the DM also learns from unverified cases and updates the DM's belief based on the machine's (and the DM's own) signal when the DM does not act. Because of salience effects and inattentional blindness, however, the DM assigns relatively less weight to this unverified information compared with information based on a verified case, for which the true state is revealed.

Formally, we consider inattentional blindness parameter $\varepsilon \in [0,1]$, such that the DM's belief b_{t-1} is updated to b_t as follows:

$$b_t = \begin{cases} \left[1 + \dfrac{\overline{b}_{t-1}}{b_{t-1}} \left(\dfrac{\mathbb{P}^{\mathsf{W}}(S_t^{\mathsf{M}} = s^{\mathsf{M}}, S_t^{\mathsf{H}} = s^{\mathsf{H}})}{\mathbb{P}^{\mathsf{B}}(S_t^{\mathsf{M}} = s^{\mathsf{M}}, S_t^{\mathsf{H}} = s^{\mathsf{H}})} \right)^{\varepsilon} \right]^{-1} \\ \text{if } \mathbb{P}(\Theta_t = \mathsf{A} \mid S_t^{\mathsf{H}} = s^{\mathsf{H}}, S_t^{\mathsf{M}} = s^{\mathsf{M}}, b_{t-1}) < r \\ \left[1 + \dfrac{\overline{b}_{t-1}}{b_{t-1}} \dfrac{\mathbb{P}^{\mathsf{W}}(S_t^{\mathsf{M}} = s^{\mathsf{M}} \mid \Theta_t = \theta)}{\mathbb{P}^{\mathsf{B}}(S_t^{\mathsf{M}} = s^{\mathsf{M}} \mid \Theta_t = \theta)} \right]^{-1} \\ \text{if } \mathbb{P}(\Theta_t = \mathsf{A} \mid S_t^{\mathsf{H}} = s^{\mathsf{H}}, S_t^{\mathsf{M}} = s^{\mathsf{M}}, b_{t-1}) \geq r. \end{cases} \tag{15}$$

Here, ε represents how less salient unverified information is compared with verified information.[7] The higher the value of ε is, the more sensitive the DM is to the informativeness of the machine's signal for an unverified case compared with a verified one. The verification bias is fully relaxed, and proper learning occurs when $\varepsilon = 1$ per Proposition 1 in Online Appendix E.[8] By contrast, our

main setup corresponds to $\varepsilon = 0$. The next theorem shows that our results continue to hold when ε is positive but sufficiently low.

Theorem 8. *Unique thresholds ε^B and ε^W exist such that*

- *When the machine is better ($\Gamma = B$), if $\varepsilon \leq \varepsilon^B$ and $p < p^B$, then b_t oscillates and is recurrent; otherwise, $b_t \xrightarrow{a.s.} 1$.*
- *When the machine is worse ($\Gamma = W$), if $\varepsilon < \varepsilon^W$ and $p > p^W$, then $b_t \xrightarrow{a.s.} X$, where X is a Bernoulli random variable; otherwise, $b_t \xrightarrow{a.s.} 0$.*

Thresholds ε^B and ε^W are, respectively, defined in (83) *and* (84) *in Online Appendix E, and p^B and p^W are in Theorem 2.*

Theorem 8 corresponds to Theorems 3 and 4 with the additional condition that ε is less than ε^Γ for $\Gamma \in \{B, W\}$, respectively. In particular, when the unverified cases are sufficiently less salient than the verified ones with $\varepsilon < \min(\varepsilon^B, \varepsilon^W)$, our main results always hold.

10. Conclusion

This paper proposes a framework in which a machine performs repeated decision tasks under the supervision of a DM. In this setup, we fully characterize the evolution of the DM's belief about the machine and overruling decisions over time. We find that mislearning can take two radically different forms: a constant change of mind (oscillation of the DM's belief per Theorem 3) and a chance of being persuaded that the machine has the wrong accuracy levels (convergence of the belief to a Bernoulli variable per Theorem 4). This contrasts with the convergence of the DM's belief to an interior point in (0, 1), which is often found in the dynamic learning literature (see e.g., confounding beliefs in Harrison et al. 2012). This analysis also provides a novel explanation for the joint production of decisions by machines and experts and suggests several guidelines for adopting or abandoning a machine.

The different forms of mislearning we uncover in this paper stem from the interaction between the DM's belief in the machine and the DM's decision to act, which, in turn, determines the DM's sampling of correct and incorrect machine predictions. The belief and resulting sampling bias interact in a negative feedback loop when the machine is better, whereas the feedback loop is positive when the machine is worse.

These learning failures do not arise from an intrinsic mistrust bias against machine-based predictions, such as algorithmic aversion. Rather, they stem from the problem of learning about a machine while actually using its predictions to make high-stakes decisions. We capture the key features of this problem with four fundamentals: informativeness, substitution, verification bias and exploration-free decisions.

Of these four, the last two conditions are crucial for our findings. Indeed, the DM always properly learns the true nature of the machine when the verification bias is sufficiently relaxed (per Theorem 8). Similarly, our no-interaction benchmark corresponds to a partial relaxation of the exploration-free condition, which also induces proper learning (see Theorem 1). In contrast, we find that the DM sometimes randomly fails to learn the machine's accuracy when its predictions complement the DM's judgment (see Theorem 7). We further expect mislearning to occur even when some of the signals are not informative although the problem can become degenerate in this case (when none of the signals are informative, for instance).

We also restrict our analysis to two possible machine types, mostly for simplicity, but our framework can be extended to account for more, possibly continuous types. Our results should not change overall as long as the previous fundamentals hold. Indeed, the DM's belief that the machine outperforms the DM's expertise is what fundamentally matters when deciding to override the machine. This, in essence, divides the different possible machine types into two distinct partitions depending on whether the type is better than the DM. In this sense, we retrieve a setup with two—albeit more convoluted—machine types.

Even though we assume them away, a DM may nonetheless be subject to mistrust biases against the machine in our setup. Our results indicate that these biases can interact with our results in a significant way. In particular, the presence of mistrust bias akin to algorithm aversion sometimes randomizes the DM's ability to properly learn the true nature of the machine. These results also provide novel hypotheses that future experimental research can test.

We focus on mistrust biases in this paper, but our framework can potentially accommodate other types of biases, such as overconfidence and loss aversion (Benjamin 2019). Further, our framework can potentially account for situations in which the DM does not perfectly know the DM's own accuracy or has a misspecified representation of the machine (Fudenberg et al. 2017). Alternatively, the machine may provide partial explanations for the machine's prescription, which may help the DM to learn the true machine accuracy (see, e.g., Puranam and Tsetlin 2021 for a way to model explainability).

Note finally that our framework may also be applied to situations in which an expert supervises another expert instead of a machine. Doing so, however, requires assuming that experts learn the level of expertise solely by observing the ex post accuracy of someone's judgments. Whereas this precise setting may exist, experts such as radiologists typically provide a rationale or causal explanation to justify their prescriptions. These explanations are also indicative of someone's knowledge and expertise. In other words, a human expert can more directly and a priori assess the quality of someone's

judgment in a way that is difficult with an ML algorithm (see, e.g., Cukier et al. 2022 for more on the difference between machine-based predictions and human cognition). In this sense, our framework is better suited for and offers a fruitful approach to exploring the issue of learning whether human expertise should overrule machine-based prescriptions.

Acknowledgments

The authors are grateful to Santiago R. Balseiro; Jean Pauphilet; and seminar attendees at Yale University, Dartmouth College, HEC Paris, The Catholic University of Portugal, Machine Learning Approaches for Finance and Management conference at Humboldt University of Berlin, Bilkent University, and the European Decision Science seminar for their valuable comments. Furthermore, the authors thank the department editor, associate editor, and anonymous reviewers for insightful comments and suggestions that greatly improved this paper. In memory of Denis Gromb.

Endnotes

[1] To be more precise, the verification event in the no-interaction benchmark is due to the DM's own judgment, and the event that the machine prescribes to act is independent of the DM's judgment conditional on the task's type.

[2] For instance, our main setup does not satisfy the missing-at-random assumption used by Begg and Greenes (1983) or the restrictions imposed on the data-generating process proposed by Zhou (1993). In addition, our no-overriding benchmark corresponds to so-called extreme verification bias (Pepe 2003) for which the estimation of accuracy parameters is impossible (Broemeling 2011).

[3] This assumption guarantees that the DM's belief in a better machine decreases (increases) upon observing an incorrect (correct) machine prediction. In contrast, assuming $\alpha^{\mathsf{W}} > \alpha^{\mathsf{B}}$ ($\beta^{\mathsf{W}} > \beta^{\mathsf{B}}$) implies that the DM's belief that the machine is better actually increases after observing a false negative (false positive) error.

[4] In particular, we have $\mathbb{P}(\Theta = \mathsf{A} | S^{\mathsf{H}} = s^{\mathsf{H}}, S^{\mathsf{M}} = s^{\mathsf{M}}, 1) = \mathbb{P}^{\mathsf{B}}(\Theta = \mathsf{A} | S^{\mathsf{H}} = s^{\mathsf{H}}, S^{\mathsf{M}} = s^{\mathsf{M}})$ and $\mathbb{P}(\Theta = \mathsf{A} | S^{\mathsf{H}} = s^{\mathsf{H}}, S^{\mathsf{M}} = s^{\mathsf{M}}, 0) = \mathbb{P}^{\mathsf{W}}(\Theta = \mathsf{A} | S^{\mathsf{H}} = s^{\mathsf{H}}, S^{\mathsf{M}} = s^{\mathsf{M}})$.

[5] See Alizamir et al. (2013), for instance, for a microfoundation of threshold r.

[6] The divergence of L_t is due to the strong law of large numbers; see Gut (2009, Theorem 8.3) for more details.

[7] To see this, consider a setup with two different absolute weights for the verified and unverified cases, say ω_v and ω_u, respectively. This setup is equivalent to the one in Section 9 by taking $\varepsilon = \omega_u/\omega_v$.

[8] This proposition is consistent with the frequentist consistency of Bayesian updating (see, e.g., Diaconis and Freedman 1986), which implies perfect learning when the verification bias is fully relaxed with $\varepsilon = 1$.

References

Acemoglu D, Dahleh MA, Lobel I, Ozdaglar A (2011) Bayesian learning in social networks. *Rev. Econom. Stud.* 78(4):1201–1236.

Ahsen ME, Ayvaci MUS, Raghunathan S (2019) When algorithmic predictions use human-generated data: A bias-aware classification algorithm for breast cancer diagnosis. *Inform. Systems Res.* 30(1):97–116.

Alizamir S, de Véricourt F, Sun P (2013) Diagnostic accuracy under congestion. *Management Sci.* 59(1):157–171.

Allon G, Drakopoulos K, Manshadi V (2021) Information inundation on platforms and implications. *Oper. Res.* 69(6):1784–1804.

Bastani H, Bastani O, Sinchaisri WP (2021a) Improving human decision-making with machine learning. Working paper, The Wharton School, Operations Information and Decisions, Philadelphia. https://parksinchaisri.github.io/files/tips.pdf.

Bastani H, Bayati M, Khosravi K (2021b) Mostly exploration-free algorithms for contextual bandits. *Management Sci.* 67(3):1329–1349.

Bates AS, Margolis PA, Evans AT (1993) Verification bias in pediatric studies evaluating diagnostic tests. *J. Pediatrics* 122(4):585–590.

Baumeister RF, Bratslavsky E, Finkenauer C, Vohs KD (2001) Bad is stronger than good. *Rev. General Psych.* 5(4):323–370.

Begg CB, Greenes RA (1983) Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics* 39(1):207–215.

Benjamin DJ (2019) Errors in probabilistic reasoning and judgment biases, Chapter 2. Douglas B, DellaVigna S, Laibson D, eds. *Handbook of Behavioral Economics: Applications and Foundations 1*, vol. 2 (North-Holland, Amsterdam), 69–186.

Besbes O, Zeevi A (2009) Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Oper. Res.* 57(6):1407–1420.

Bordalo P, Gennaioli N, Shleifer A (2012) Salience theory of choice under risk. *Quart. J. Econom.* 127(3):1243–1285.

Boyacı T, Özer Ö (2010) Information acquisition for capacity planning via pricing and advance selling: When to stop and act? *Oper. Res.* 58(5):1328–1349.

Boyacı T, Canyakmaz C, de Véricourt F (2023) Human and machine: The impact of machine input on decision-making under cognitive limitations. *Management Sci.*, ePub ahead of print March 31, https://doi.org/10.1287/mnsc.2023.4744.

Broemeling LD (2011) Bayesian estimation of combined accuracy for tests with verification bias. *Diagnostics (Basel)* 1(1):53–76.

Camacho N, Donkers B, Stremersch S (2011) Predictably non-Bayesian: Quantifying salience effects in physician learning about drug quality. *Marketing Sci.* 30(2):305–320.

Cheung WC, Simchi-Levi D, Wang H (2017) Dynamic pricing and demand learning with limited price experimentation. *Oper. Res.* 65(6):1722–1731.

Coutts A (2019) Good news and bad news are still news: Experimental evidence on belief updating. *Experiment. Econom.* 22(2):369–395.

Cowgill B (2019) Bias and productivity in humans and machines. Research paper, Columbia Business School, New York.

Cukier K, Mayer-Schönberger V, de Véricourt F (2022) *Framers: Human Advantage in an Age of Technology and Turmoil* (Dutton-Penguin Random House, New York).

Diaconis P, Freedman D (1986) On the consistency of Bayes estimates. *Ann. Statist.* 14(1):1–26.

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Experiment. Psych. General* 144(1):114–126.

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Sci.* 64(3):1155–1170.

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

Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, Coughlin JF, Guttag JV, Colak E, Ghassemi M (2021) Do as AI say: Susceptibility in deployment of clinical decision-aids. *NPJ Digital Medicine* 4(1):1–8.

Grady D (2019) AI took a test to detect lung cancer. It got an A. *The New York Times* (May 20), 20.

Greenes RA, Begg CB (1985) Assessment of diagnostic technologies: Methodology for unbiased estimation from samples of selectively verified patients. *Investigative Radiology* 20(7):751–756.

Guo Y, Zhang C, Yang XJ (2020) Modeling trust dynamics in human-robot teaming: A Bayesian inference approach. *Extended Abstracts*

2020 CHI Conf. Human Factors Comput. Systems (Association for Computing Machinery, New York), 1–7.

Gut A (2009) *Stopped Random Walks* (Springer, New York).

Harrison JM, Keskin NB, Zeevi A (2012) Bayesian dynamic pricing policies: Learning and earning under a binary prior distribution. *Management Sci.* 58(3):570–586.

Herrera H, Hörner J (2013) Biased social learning. *Games Econom. Behav.* 80:131–146.

Hujoel IA, Jansson-Knodell CL, Hujoel PP, Hujoel ML, Choung RS, Murray JA, Rubio-Tapia A (2021) Estimating the impact of verification bias on celiac disease testing. *J. Clinical Gastroenterology* 55(4):327–334.

Ibrahim R, Kim S-H, Tong J (2021) Eliciting human judgment for prediction algorithms. *Management Sci.* 67(4):2314–2325.

Kahneman D (1973) *Attention and Effort*, vol. 1063 (Prentice-Hall, Hoboken, NJ).

Keskin NB, Birge JR (2019) Dynamic selling mechanisms for product differentiation and learning. *Oper. Res.* 67(4):1069–1089.

Kubat M (2017) *An Introduction to Machine Learning*, vol. 2 (Springer, Cham, Switzerland).

Lebovitz S, Levina N, Lifshitz-Assaf H (2021) Is AI ground truth really "true"? The dangers of training and evaluating AI tools based on experts' know-what. *Management Inform. Systems Quart.* 45(3):1501–1525.

Lebovitz S, Lifshitz-Assaf H, Levina N (2022) To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organ. Sci.* 33(1):126–148.

Lee HCB, Ba S, Li X, Stallaert J (2018) Salience bias in crowdsourcing contests. *Inform. Systems Res.* 29(2):401–418.

McKendrick J (2021) AI adoption skyrocketed over the last 18 months. Accessed February 18, 2022, https://hbr.org/2021/09/ai-adoption-skyrocketed-over-the-last-18-months.

Möbius MM, Niederle M, Niehaus P, Rosenblat TS (2022) Managing self-confidence: Theory and experimental evidence. *Management Sci.* 68(11):7793–7817.

Özer Ö, Zheng Y (2018) Trust and trustworthiness. Donohue K, Katok E, Leider S, eds. *The Handbook of Behavioral Operations* (Wiley, Hoboken, NJ), 489–523.

Pepe MS (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction* (Oxford University Press, New York).

Petscavage JM, Richardson ML, Carr RB (2011) Verification bias: An underrecognized source of error in assessing the efficacy of medical imaging. *Acad. Radiology* 18(3):343–346.

Puranam P, Tsetlin I (2021) The limits to explainability. Working paper, INSEAD, Singapore. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3914597.

Ransohoff DF, Feinstein AR (1978) Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England J. Medicine* 299(17):926–930.

Reardon S (2019) Rise of robot radiologists. *Nature* 576(7787):S54–S58.

Schwartzstein J (2014) Selective attention and learning. *J. Eur. Econom. Assoc.* 12(6):1423–1452.

Simon HA (1955) A behavioral model of rational choice. *Quart. J. Econom.* 69(1):99–118.

Smith L, Sørensen P (2000) Pathological outcomes of observational learning. *Econometrica* 68(2):371–398.

Soll JB, Mannes AE (2011) Judgmental aggregation strategies depend on whether the self is involved. *Internat. J. Forecasting* 27(1):81–102.

Stone M (1961) The opinion pool. *Ann. Math. Statist.* 32(4):1339–1342.

Sun J, Zhang DJ, Hu H, Van Mieghem JA (2021) Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Sci.* 68(2): 846–865.

Taylor SE, Thompson SC (1982) Stalking the elusive "vividness" effect. *Psych. Rev.* 89(2):155–181.

Tiefenbeck V, Goette L, Degen K, Tasic V, Fleisch E, Lalive R, Staake T (2018) Overcoming salience bias: How real-time feedback fosters resource conservation. *Management Sci.* 64(3):1458–1476.

Tschandl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H, Gutman D, et al. (2019) Comparison of the accuracy of human readers vs. machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study. *Lancet Oncology* 20(7):938–947.

Van Donselaar KH, Gaur V, Van Woensel T, Broekmeulen RA, Fransoo JC (2010) Ordering behavior in retail stores and implications for automated replenishment. *Management Sci.* 56(5): 766–784.

Wang C, Zhang C, Yang XJ (2018) Automation reliability and trust: A Bayesian inference approach. *Proc. Human Factors Ergonomics Soc. Annual Meeting*, vol. 62 (Sage Publications, Los Angeles), 202–206.

Whiting PF, Rutjes AW, Westwood ME, Mallett S, QUADAS-2 Steering Group (2013) A systematic review classifies sources of bias and variation in diagnostic test accuracy studies. *J. Clinical Epidemiology* 66(10):1093–1104.

Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, Jung K, et al. (2019) Do no harm: A roadmap for responsible machine learning for healthcare. *Nature Medicine* 25(9):1337–1340.

Zhou X-H (1993) Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. *Comm. Statist. Theory Methods* 22(11):3177–3198.