de Véricourt, Francis; Gurkan, Huseyin

Working Paper

Is your machine better than you? You may never know

ESMT Working Paper, No. 22-02

Provided in Cooperation with:
ESMT European School of Management and Technology, Berlin

Suggested Citation: de Véricourt, Francis; Gurkan, Huseyin (2022) : Is your machine better than you? You may never know, ESMT Working Paper, No. 22-02, European School of Management and Technology (ESMT), Berlin

This Version is available at:
https://hdl.handle.net/10419/259795

ISSN 1866-3494

May 23, 2022

ESMT Working Paper 22-02

Is your machine better than you? You may never know

Francis de Véricourt, ESMT European School of Management and Technology

Huseyin Gurkan, ESMT European School of Management and Technology

Is Your Machine Better Than You? You May Never Know.

Francis de Véricourt, Huseyin Gurkan*
ESMT Berlin, {francis.devericourt, huseyin.gurkan}@esmt.org

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-stake decisions can properly assess whether the machine produces better recommendations. To that end, we study a set-up, 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, she updates her belief about the machine. However, the DM observes the machine's correctness only if she ultimately decides to act on the task. Further, the DM sometimes overrides the machine depending on her belief, which affects learning. In this set-up, 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, i.e., she 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. Our results highlight some fundamental limitations in determining whether machines make better decisions than experts and provide a novel explanation for human-machine complementarity.

Key words: 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 have demonstrated 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, has been limited (Wiens et al. 2019) and challenged by the tendency of decision makers to override algorithmic prescriptions. For instance, Sun et al. (2021) find in warehouse operations that employees significantly deviated from the recommendations of an algorithm. Lebovitz et al. (2022) also report how a team of

* The authors are grateful to Santiago R. Balseiro, Denis Gromb, Jean Pauphilet and the seminar attendees at Yale University, Dartmouth College, HEC Paris, The Catholic University of Portugal, ML Approaches for Finance and Management conference at Humboldt University of Berlin, Bilkent University and the European Decision Science seminar for their valuable comments.

radiologists in a large US-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, sometimes 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. A perhaps more fundamental question remains: to which extent the nature of the decision problem and its context prevent the decision maker from learning whether a machine produces better prescriptions.

In this paper, we explore fundamental limitations in a decision maker's ability to properly learn whether a machine is superior to human expertise. Importantly, these limitations do not stem from a mistrust bias against the machine. We focus on situations where 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 datasets, however, never fully capture the ground truth, and the issue of empirical generalizability remains (see, e.g., Lebovitz et al. 2021). Thus, an expert may continue to observe and adjust her belief about the machine after adopting it. However, because the machine is deployed and, hence, makes prescriptions with real consequences, learning can be impaired in ways that do not exist during the training phase of the algorithm.

In particular, the correctness of a machine's prediction is not always observable once the machine is deployed. Further, the decision maker cannot always experiment with real cases to learn more about the machine's actual accuracy. This is especially true for high-stake decisions found, for instance, in the medical and judiciary sectors. Our goal is thus to study how these limitations may induce the expert to mislearn whether a machine is superior to her own judgment.

To that end, we analyze a set-up, in which a machine performs repeated decision tasks under the supervision of a decision maker (DM). Each task consists in deciding whether or not 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 her own expertise. Crucially, the DM is uncertain about whether the machine makes better or worse decisions than she does, but as the DM observes the correctness of the different machine's prescriptions, she forms a belief about the machine's true accuracy.

We consider situations in which the DM observes the correctness of the machine's prediction and updates her belief accordingly only if the action is actually taken (e.g., when a biopsy is performed). In other words, the DM is subject to a form of selection bias referred to as *verification bias* in the statistic literature, such that the results of diagnostic tests determine whether the true state of the world (the presence or absence of disease) is verified (e.g., Pepe 2003, p. 169). Further, the DM

only decides what is best for the task at hand and thus never acts for the purpose of observing the machine's accuracy. In this sense, the DM's decisions are *exploration-free*. This restriction may be for legal or ethical concerns, as in the medical and judiciary sectors (Bastani, Bayati and Khosravi 2021).

In addition to these limitations, we consider certain features of the machine's and DM's predictions. Specifically, the DM is able to decide solely based on either her own judgment or the machine's prescription. In this sense, the DM has adequate expertise and the machine is sufficiently accurate to make a prediction for the task, i.e., both the machine and the DM generate *informative* signals. Further, the machine and the DM are *substitutes* in that the DM's accuracy is either better or worse than the machine's.

Our framework can account for situations where the machine and DM may complement one another, as discussed in Section 8. Nonetheless, we focus in this paper on substitution for two reasons. First, our goal is to propose alternative explanations for overriding machines to algorithm aversion, the studies of which assume substitution (Dietvorst et al. 2015). 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.

In this set-up, we study the evolution of the DM's belief and overruling decisions over time. This enables identification of the situations in which the DM properly learns whether the machine makes better predictions than her. 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. Note that studying the asymptotic behavior of believes as we do differs from most research problems found in the operations literature on learning (see Section 2). These studies typically consider learning problems with exploration, which eliminates the issue of the belief's asymptotic behavior since the DM always properly learns in the limit. The question becomes then how fast and efficient this learning can be. In contrast, our setting is exploration free, which mutes the problem of efficient learning, but raises the question of whether the DM properly learns in the limit.

To study the DM's learning behavior over time, we first control for the role that overriding decisions has in the DM's ability to learn. To that end, we analyze a benchmark (Section 4) in which the DM never overrides the machine and always follows its prescription. The DM, however, observes the machine and forms a belief about its true nature. In this set-up, we find that the DM's ability to learn depends on the DM's prior about the task. This prior corresponds to the probability that an action is required for the task at hand, before the machine produces a prescription or the DM forms a judgment on the task.

In this benchmark, the DM properly learns that the machine is better (resp. worse), if the prior about the task is above (resp. below) a certain threshold. In other words, the DM can end up believing that the machine makes worse predictions than her, even though the machine is actually better. This happens 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.

The key driver for these learning failures stems from the verification bias. Without overriding, this feature implies that the DM learns about the machine only when it prescribes to act. With overriding, however, the DM sometimes learns about the machine when it prescribes not to act. This happens each time the DM decides to act against the machine's prescription. Hence, overriding promotes learning and may thus offset the learning failures we observe in the benchmark.

To explore this further, we turn to our main set-up where the DM can overrule the machine. This occurs when the DM's judgment contradicts the machine's prescription and the DM sufficiently believes that the machine is worse. In this setting, we find that the prior about the task plays the same role as in the benchmark and fully determines when the DM properly learns. When the DM fails to learn, however, the overriding decisions actually change the nature of mislearning compared to that in the benchmark.

Specifically, 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 the machine's prescriptions. Further, and perhaps more interestingly, the DM sometimes treats the machine as if its prediction complements her own judgment, while in fact, the two are full substitutes.

In addition, when the machine is worse and the DM's prior about the task is high, the DM's belief actually 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 sometimes ends up believing that the machine is better when it is actually worse.

Taken together, these results uncover several limitations in a decision-maker's ability to learn whether a machine should be overruled. They further provide a novel rationale—the uncertainty about the machine's true performance—for why human experts may co-produce their decisions with a machine. Our findings also enable experts to anticipate how their beliefs about the machine may evolve over time. This, in turn, provides guidelines for inferring whether a machine makes better or worse predictions than experts, once it is deployed (see Section 6). Importantly, the mislearning behaviors we characterize do not stem from an intrinsic algorithm aversion but rather from the

problem's structure, which we capture with four fundamentals: verification bias, exploration-free decisions, informativeness and substitution.

Nonetheless, a DM who faces this problem's structure 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) by altering our main set-up to account for biases in the decision to act (using the opinion aggregation procedure proposed by Stone 1961) and in the Bayesian updating process (by introducing a negativity bias - see Baumeister et al. 2001). We find that downplaying the machine's prescription when deciding to act does not affect the structure of our results. However, our results change when the DM disproportionately drops her belief upon observing a machine's failure. In particular, the DM does not always properly learn that the machine is better when the posterior about the task is sufficiently high, as in the main set-up. Instead, the DM's belief can converge to a Bernoulli random variable, a phenomena that occurs only if the machine is actually worse in the absence of mistrust bias. In this sense, algorithm aversion sometimes interacts with our four fundamentals (verification bias in particular) to randomize the DM's ability to learn.

After reviewing the literature in Section 2, we present the model in Section 3. In Section 4, we analyze the no-overriding benchmark and then focus on the main set-up 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. 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 co-production of decisions by a machine and a DM may improve performance. For instance, Boyaci et al. (2020) 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 Boyaci et al. 2020 for additional references on formal models of machine-human interactions). Machine learning algorithms have also been proposed to provide interpretable cues to help decision makers improve their decisions (see Bastani, Bastani and Sinchaisri 2021, for instance). This stream of research

further explores how to use human judgment to train or improve an algorithm (Van Donselaar et al. 2010, Ibrahim et al. 2021, Cowgill 2019).

We contribute to this literature by providing a novel rationale for why a DM may treat the machine's prescriptions as a complement to her judgment In fact, this stream of research typically assumes that the machine's accuracy is known and complements the DM's judgement. 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 found in an experimental set-up 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) observed that packing workers at the warehouses of the Alibaba Group regularly deviated from algorithmic prescriptions, which reduced operational efficiency. Several approaches have been explored to reduce deviations such as these, either with field experiments (Sun et al. 2021) or in the lab (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 set-up 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). 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 set-up, 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.

This verification bias is a form of selection bias, which was first introduced by Ransohoff and Feinstein (1978) in the context of diagnostic test accuracy. This notion has been extensively studied, especially in the biostatistics and medical literature. Most of this research focuses on developing estimators based on the maximum likelihood to correct bias. In contrast, our study concerns asymptotic behavior in a Bayesian framework. More importantly, the assumptions required to correct for

this bias do not hold in our set-up. For instance, our main set-up with overriding 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 benchmark model corresponds to so-called extreme verification bias (Pepe 2003, p. 180), for which the estimation of accuracy parameters is impossible (Broemeling 2011).

Finally, our work is related to the vast literature on learning problems, which have been extensively studied in management science and operations management. For instance, studies have considered price experimentation to learn demand curves by focusing on the tradeoffs 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 where exploring is not possible. Thus, the DM optimizes within each period and mislearning may emerge in our set-up.

In this sense, our approach resembles Harrison et al. (2012) which analyzes myopic pricing policies (see Section 4 in particular). In their set-up, 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 set-up, 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 Smith and Sørensen 2000, Acemoglu et al. 2011, 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) analyzes a set-up with short-lived myopic investors, in which only investing decisions are observable. Although this may resemble our set-up, 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 set-up (see Propositions 1 and 4 in Herrera and Hörner 2013), while it may not converge in ours.

3. Model Description

We consider a decision maker (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 her own judgment. As the accuracy of the machine's predictions is revealed over time, the DM forms a belief about whether or not she 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 in deciding whether or not 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 she should act.

To perform this task, the DM applies her expertise and elicits imperfect signal $S^{\mathsf{H}} \in \{+, -\}$, such that $S^{\mathsf{H}} = +$ (resp., $S^{\mathsf{H}} = -$) indicates that $\Theta = \mathsf{A}$ (resp., $\Theta = \mathsf{NA}$). 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 her own. Specifically, we denote the machine's type as $\Gamma \in \{\mathsf{B}, \mathsf{W}\}$. When $\Gamma = \mathsf{B}$ (resp., $\Gamma = \mathsf{W}$), signal S^{M} is *better* (resp., *worse*) than signal S^{H}, and the sensitivity and specificity of the signal are equal to $(\alpha^{\mathsf{B}}, \beta^{\mathsf{B}})$ (resp., $(\alpha^{\mathsf{W}}, \beta^{\mathsf{W}})$). The machine is better (resp., worse) in the sense that the DM never (resp., 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 where $\alpha^{\mathsf{B}} > \alpha^{\mathsf{W}}$ and $\beta^{\mathsf{B}} > \beta^{\mathsf{W}}$.[1] This is only for the sake of clarity, as all of our results extend to the more general case.

Probability $b \triangleq \mathbb{P}\big(\Gamma = \mathsf{B}\big)$ denotes then the DM's prior belief that the machine outperforms her 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 $\mathbb{P}^{\Gamma}(S^{\mathsf{M}} = -, \Theta = \mathsf{NA}) = \beta^{\Gamma} \bar{p}$ for $\Gamma \in \{\mathsf{B}, \mathsf{W}\}$ (with $\bar{x} = 1 - x$ for $x \in [0, 1]$).

Based on realizations s^{H} and s^{M} of signals S^{H} and S^{M}, respectively, and her belief b about the machine, the DM updates her 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).[2]

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

[2] 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 $\mathbb{P}(\Theta = \mathsf{A} \,|\, S^{\mathsf{H}} = s^{\mathsf{H}}, S^{\mathsf{M}} = s^{\mathsf{M}}, 0) = \mathbb{P}^{\mathsf{W}}(\Theta = \mathsf{A} \,|\, S^{\mathsf{H}} = s^{\mathsf{H}}, S^{\mathsf{M}} = s^{\mathsf{M}})$.

The DM then decides to act if and only if her posterior belief is above a positive threshold r, i.e., $\mathbb{P}(\Theta = \mathsf{A} \,|\, S^{\mathsf{H}} = s^{\mathsf{H}}, S^{\mathsf{M}} = s^{\mathsf{M}}, b) \geq r$; 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.[3]

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 or not the posterior belief is larger than threshold r, i.e., 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, i.e., when $s^{\mathsf{H}} = s^{\mathsf{M}}$. One signal is then sufficient for the DM to decide since 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 where the machine and the DM are full substitutes in the following sense.

Substitution: We assume that a type B machine always overrides the DM's judgment, while 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 or not 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.

[3] See, Alizamir et al. (2013) for instance, for a micro foundation of threshold r.

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 where 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 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.

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 she 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 her belief b_{t-1} to posterior b_t according to Bayes' rule, if the DM observes type Θ_t.

Verification Bias: The DM, however, observes the task's type and updates her belief accordingly if and only if an action is taken. Because decisions are exploration-free, the verification bias implies that the DM updates her 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,

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

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 observe 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, she increases (resp., decreases) her belief that the machine is better if the machine's prescription turns out to be correct (resp., wrong).

Overall, this set-up describes a process through which the DM focuses her attention on making a series of diagnostic decisions (e.g., whether to run different biopsies) based on her expertise and the machine, until the observation of the correctness of the machine's prediction prompts her to revise her belief about the machine.

When this happens, 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 her belief converges over time to 1 ($b_t \xrightarrow{\text{a.s.}} 1$) when the machine is better ($\Gamma = \mathsf{B}$) and converges to 0 ($b_t \xrightarrow{\text{a.s.}} 0$) when the machine is worse ($\Gamma = \mathsf{W}$). (Notation $\xrightarrow{\text{a.s.}}$ indicates almost-sure convergence.) In contrast, the DM mislearns the machine's type when $b_t \xrightarrow{\text{a.s.}} 0$ (resp., 1) and $\Gamma = \mathsf{B}$ (resp., $\Gamma = \mathsf{W}$). 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 $\tau > 0$, $\mathbb{P}(Y_t \in \mathcal{I} \text{ for some } t > \tau \,|\, Y_\tau \in \mathcal{I}) = 1$ (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. No-Overriding Benchmark

We first study the setting where the DM never overrides the machine's prediction but continues to form a belief about the type of the machine. Studying this benchmark enables us to identify in the next section, the effect of the DM's own decisions on her ability to learn the machine's type. Without overriding, prior b_{t-1} does not determine whether an action is taken and, thus, whether the task's type is observed ex post. Hence, the second mechanism by which prior b_{t-1} influences posterior b_t is mute in this benchmark. Learning occurs through only the first mechanism, i.e., the application of Bayes' rule when the machine prescribes to act.

More specifically, the DM acts if and only if $S_t^{\mathsf{M}} = +$ regardless of her own judgment S_t^{H}. The condition for the DM to act, $\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^{\mathsf{H}}, S_t^{\mathsf{M}}, b_{t-1}) \geq r$, thus reduces to $\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^{\mathsf{M}}, b_{t-1}) \geq r$, which is equivalent to $S_t^{\mathsf{M}} = +$ for any b_{t-1} due to Informativeness (2)-(3). Equation (6) then becomes

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

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 $L_t \triangleq \log\left(\frac{b_t}{1-b_t}\right)$, such that

$$L_t = L_{t-1} + R_t^{\mathsf{M}},$$

where $(R_t^{\mathsf{M}})_{t\in\mathbb{N}}$ are 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 resp. $L_t \xrightarrow{\text{a.s.}} -\infty$) per the continuous mapping theorem.

Log-likelihood ratio L_t is a random walk governed by jumps $(R_t^{\mathsf{M}})_{t\in\mathbb{N}}$, which capture the magnitude and direction of the belief's updates. These random jumps take three possible values: a positive (resp., negative) value when the DM observes that the machine correctly (resp., wrongly) prescribes to act, i.e., $S_t^{\mathsf{M}} = +$ and $\Theta_t = \mathsf{A}$ (resp., $\Theta_t = \mathsf{NA}$), or zero when the task's type is not observed, i.e., $S_t^{\mathsf{M}} = -$. The asymptotic behavior of L_t is then fully determined by the sign of the mean $\mathbb{E}^{\Gamma}[R_t^{\mathsf{M}}]$. If $\mathbb{E}^{\Gamma}[R_t^{\mathsf{M}}] > 0$ (resp., < 0), then log-likelihood ratio L_t increases in expectation and converges almost surely to $+\infty$ (resp., $-\infty$),[4] while L_t does not converge when $\mathbb{E}^{\Gamma}[R_t^{\mathsf{M}}] = 0$. Based on these properties, we characterize next the DM's ability to learn the machine's type when the DM does not override its prescriptions.

THEOREM 1 **(Learning without Overriding)**. *Unique thresholds p^{B} and p^{W} exist such that $p^{\mathsf{B}} < p^{\mathsf{W}}$ and,*

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

Further, we have

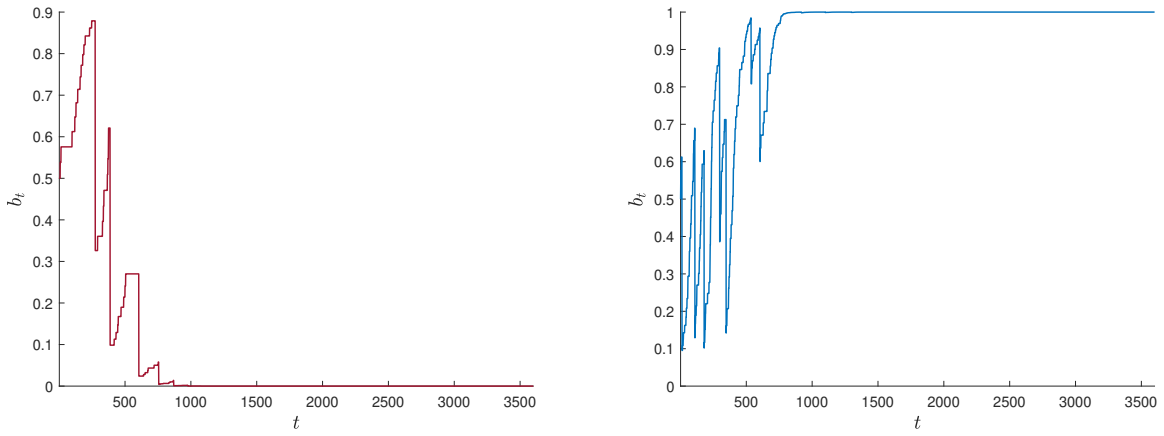
$$p^{\mathsf{B}} \triangleq \frac{\bar{\beta}^{\mathsf{B}} \log\left(\frac{\bar{\beta}^{\mathsf{W}}}{\bar{\beta}^{\mathsf{B}}}\right)}{\bar{\beta}^{\mathsf{B}} \log\left(\frac{\bar{\beta}^{\mathsf{W}}}{\bar{\beta}^{\mathsf{B}}}\right) + \alpha^{\mathsf{B}} \log\left(\frac{\alpha^{\mathsf{B}}}{\alpha^{\mathsf{W}}}\right)} \text{ and } p^{\mathsf{W}} \triangleq \frac{\bar{\beta}^{\mathsf{W}} \log\left(\frac{\bar{\beta}^{\mathsf{W}}}{\bar{\beta}^{\mathsf{B}}}\right)}{\bar{\beta}^{\mathsf{W}} \log\left(\frac{\bar{\beta}^{\mathsf{W}}}{\bar{\beta}^{\mathsf{B}}}\right) + \alpha^{\mathsf{W}} \log\left(\frac{\alpha^{\mathsf{B}}}{\alpha^{\mathsf{W}}}\right)}.$$

To prove this result, we first establish that $\mathbb{E}^{\Gamma}[R_t^{\mathsf{M}}] > 0$ (resp. < 0) if $p > p^{\Gamma}$ (resp. $< p^{\Gamma}$) for $\Gamma \in \{\mathsf{B}, \mathsf{W}\}$, and then apply the continuous mapping theorem. (All proofs are in the appendix.) Figures 1 and 2 depict how the DM's belief evolves for a given machine's type. The figures (as all figures do in this paper) illustrate the limiting behavior of the DM's belief by plotting sample paths over time.

In essence, Theorem 1 states that the DM's ability to learn depends on her prior about the task as well as the machine's type. The DM learns that the machine is worse (resp., better) when her prior is below (resp., 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 her. Indeed, when prior p is low ($p < p^{\mathsf{B}}$), the DM learns that the machine is worse ($b_t \xrightarrow{\text{a.s.}} 0$, see Figure 1a), while the machine is actually

[4] The divergence of L_t is due to the strong law large numbers; see Gut (Theorem 8.3 in p. 68 2009) for more details.

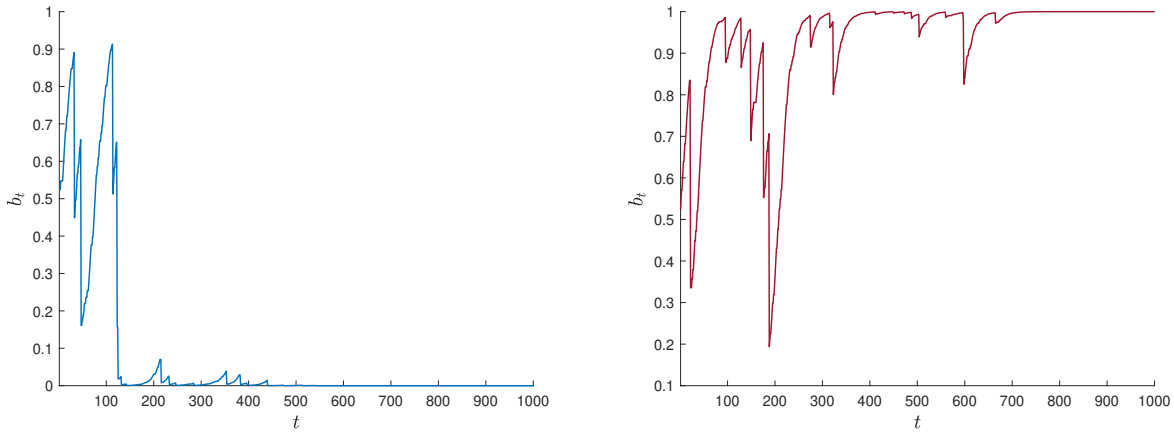
Figure 1 **The DM's belief b_t at the no-overriding benchmark when the machine is better, $\Gamma = \mathsf{B}$**



(a) $p < p^{\mathsf{B}}$ (b) $p > p^{\mathsf{B}}$

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) $p = 0.2$.

Figure 2 **The DM's belief b_t at the no-overriding benchmark when the machine is worse, $\Gamma = \mathsf{W}$**



(a) $p < p^{\mathsf{W}}$ (b) $p > p^{\mathsf{W}}$

Note. $\alpha^{\mathsf{H}} = \beta^{\mathsf{H}} = 0.95$, $\alpha^{\mathsf{B}} = \beta^{\mathsf{B}} = 0.99$, $\alpha^{\mathsf{W}} = \beta^{\mathsf{W}} = 0.9$, $p^{\mathsf{W}} = 0.72$ and $r = 0.8$, (a) $p = 0.71$, (b) $p = 0.75$.

better ($\Gamma = \mathsf{B}$). Similarly, when prior p is high ($p > p^{\mathsf{W}}$), the DM learns that the machine is better ($b_t \xrightarrow{\text{a.s.}} 1$, see Figure 2b), while the machine is actually worse ($\Gamma = \mathsf{W}$).

Theorem 1 stems from the fact that, in this benchmark, the DM acts only if the machine's signal is positive. To see this, recall first that the DM's belief increases (resp., decreases) when the DM observes a correct (resp., incorrect) machine recommendation. Because the DM is able to observe this only when the machine's signal is positive, the DM increases her belief if and only if the machine correctly prescribes to act ($S_t^{\mathsf{M}} = +$ and $\Theta_t = \mathsf{A}$) and decreases her belief if and only if the machine wrongly prescribes to act ($S_t^{\mathsf{M}} = +$ and $\Theta_t = \mathsf{NA}$). And because the belief is unchanged

when the machine's signal is negative, this further means that the DM's belief weakly increases if the task requires an action (i.e., if $\Theta_t = \mathsf{A}$), and weakly decreases otherwise (i.e., if $\Theta_t = \mathsf{NA}$).

In other words, the frequency with which the DM increases and decreases her belief is fully determined by priors p and $\bar{p}$, respectively. In particular, the DM increases her belief more frequently when the task is more likely to require an action ($\Theta_t = \mathsf{A}$), i.e., prior p takes higher values. The magnitudes of these changes in beliefs, however, do not depend on prior p but are determined by the accuracy levels 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 Set-up: Learning with Overriding

In our main set-up, the DM can overrule the machine's recommendations by exerting her own judgment. Overruling the machine actually mitigates the effects of prior p that induce mislearning in the no-overriding benchmark. Indeed, the DM can evaluate the machine's accuracy only when the machine prescribes to act in this benchmark. With overriding, however, the DM may decide to act and thus observe the task's type when the machine prescribes not to act. The key question, therefore, is to which extent overriding enables the DM to learn the true type of the machine.

More specifically, recall that due to Informativeness, the DM always decides according to her 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 depending on her current belief b_{t-1}. The following result determines when such overriding decisions occur. (The result follows from Substitution (4)-(5) and the continuity of the posterior probabilities in b_{t-1}; see Appendix B.)

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

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

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

Lemma 1 states that when the DM's judgment contradicts the machine's prescription, i.e., $S_t^{\mathsf{H}} \neq S_t^{\mathsf{M}}$, the DM overrides the machine if and only if her belief in a better machine is sufficiently low, i.e., 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 follows 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 her judgment; this is 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—her 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 she 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 her 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 2 **(Learning with Overriding)**. *When the machine is better, i.e.* $\Gamma = B$*, if* $p \leq p^B$*, then* b_t *oscillates and is recurrent; otherwise* $b_t \xrightarrow{a.s.} 1$.

Thus, with overriding, the DM's ability to learn the machine's type continues to depend on whether her 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 1). Specifically, the DM learns that the machine is indeed better ($b_t \xrightarrow{\text{a.s.}} 1$) if and only if prior p is sufficiently high with $p > p^{\mathsf{B}}$. Figure 3b illustrates this case and exhibits asymptotic behavior similar to that in Figure 1b for the no-overriding benchmark.

The DM's ability to override the machine, however, fundamentally changes the nature of mislearning. Indeed, when prior p is such that $p \leq p^{\mathsf{B}}$, the DM wrongly learns that the machine is worse in the no-overriding benchmark. With overriding, however, the belief oscillates as illustrated in Figure 3a. And 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, i.e.,* $\Gamma = \mathsf{B}$ *and* $p \leq p^{\mathsf{B}}$*, intervals* $(0, b^{\mathsf{H}}]$*,* $(b^{\mathsf{H}}, b^{\mathsf{M}})$ *and* $[b^{\mathsf{M}}, 1)$ *are recurrent for belief* b_t*.*

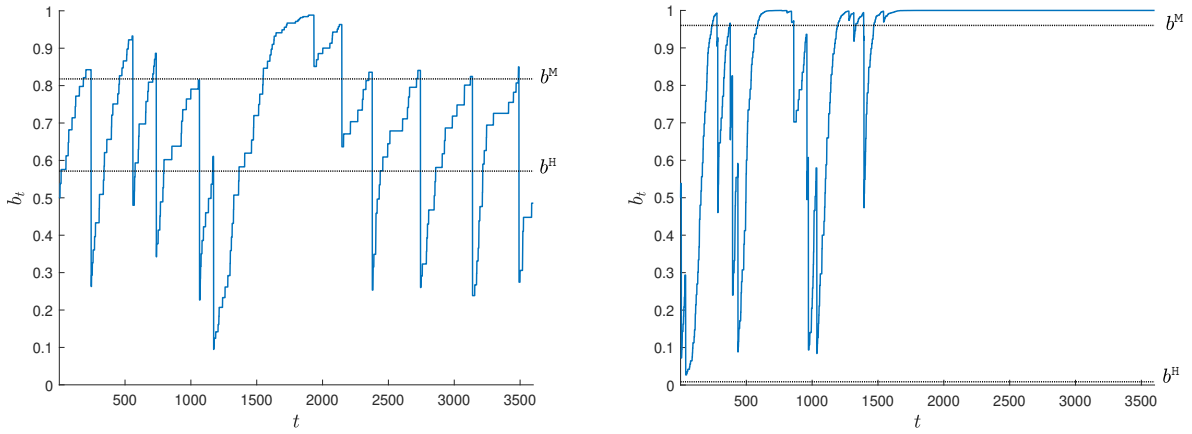
Hence, when the DM sufficiently believes that the machine is better, she never overrules it and we retrieve the dynamics of the no-overriding benchmark. That is, when $b_t > b^{\mathsf{M}}$, learning is entirely driven by whether or not a machine's prescription to act is correct. And because prior p is low, the frequency of these correct predictions is also low, so the belief is decreasing in expectation.

In contrast, when the DM sufficiently believes that the machine is worse with $b_t < b^{\mathsf{H}}$, she 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's type, and because the machine is actually better, the belief increases in expectation.

Therefore, belief b_t is pushed back downward when it reaches high values ($b_t > b^{\mathsf{M}}$), and pushed upward when it takes low values ($b_t < b^{\mathsf{H}}$). Hence, the DM never fully learns that the machine is better, but due to overriding, never mislearns that it is worse either. In this sense, the DM always remains in perpetual uncertainty about whether or not to disregard the machine.

Interestingly, this long-run uncertainty induces the DM to sometimes treat the machine's prescription as a complement to her judgment. This happens when the belief reaches $b_t \in (b^{\mathsf{H}}, b^{\mathsf{M}})$, which is a recurrent event. In this case, the DM and the machine co-produce 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.

Figure 3 **The DM's belief b_t when the machine is better, $\Gamma = \mathsf{B}$**



(a) $p \leq p^{\mathsf{B}}$ (b) $p > p^{\mathsf{B}}$

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) $p = 0.2$, $b^{\mathsf{H}} = 0.01$, $b^{\mathsf{M}} = 0.96$.

5.2. Learning When the Machine is Worse

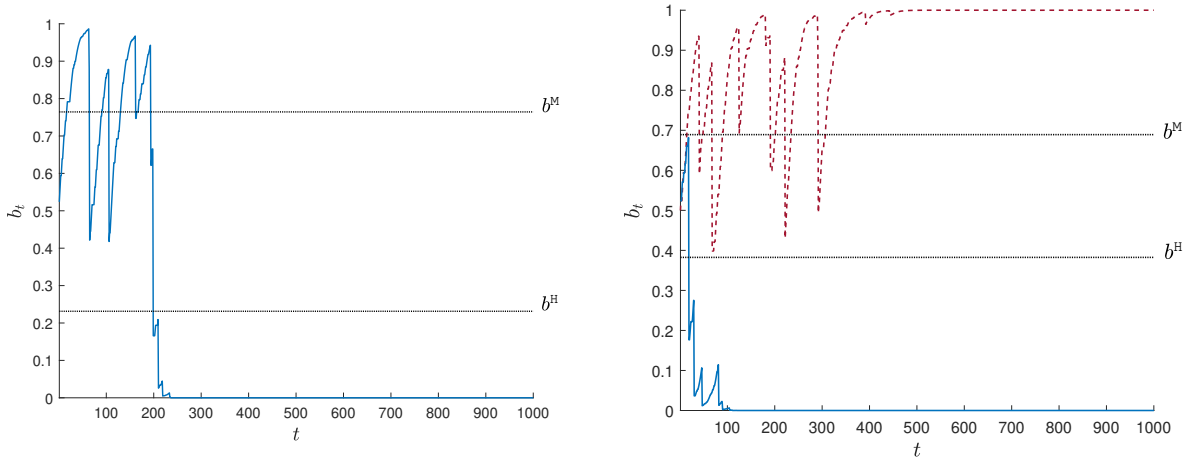
Per Theorems 1 and 2, the DM properly learns that the machine is good when prior p takes high values (i.e., $p > p^{\mathsf{B}}$), whether the DM can override the machine or not. In this case, overriding essentially prevents the DM from wrongly learning that the machine is worse, which creates a perpetual 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 3 **(Learning with Overriding)**. *When the machine is worse, i.e., $\Gamma = \mathsf{W}$, if $p \leq p^{\mathsf{W}}$, then $b_t \xrightarrow{a.s.} 0$; otherwise, $b_t \xrightarrow{a.s.} X$ where X is a Bernoulli random variable.*

Theorem 3 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 ($b \xrightarrow{\text{a.s.}} 0$) if prior p takes low values ($p < p^{\mathsf{W}}$, where threshold p^{W} is, again, the same as that in the no-overriding benchmark). Figure 4a illustrates this point, and depicts a sample path of b_t, which is similar to the one in Figure 2a for the no-overriding benchmark.

When the prior is high ($p > p^{\mathsf{W}}$), 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 nor 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, while it is actually worse. Figure 4b 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.

Figure 4 **The DM's belief b_t when the machine is worse, $\Gamma = \mathsf{W}$**



(a) $p \leq p^{\mathsf{W}}$ (b) $p > p^{\mathsf{W}}$

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

Similar to the better machine case, learning is driven by prior p as in the no-overriding benchmark when the belief is high ($b_t > b^{\mathsf{M}}$), and by the true type of the machine when the belief is low ($b_t < b^{\mathsf{H}}$). In the latter case, the belief decreases in expectation since the machine is worse.

Thus, for low prior $p < p^{\mathsf{W}}$, 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 $p > p^{\mathsf{W}}$, 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. Whether the belief reaches high or low values is determined by realizations of the different signals and the task types and is thus random. Note that when the belief takes intermediary values ($b_t \in (b^{\mathsf{H}}, b^{\mathsf{M}})$) it can either decrease or increase in expectation depending on the problem parameters. However this region is transient since the belief is pushed away from the region when the belief is more extreme ($b_t \notin (b^{\mathsf{H}}, b^{\mathsf{M}})$).

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 her own judgment, even though the two are actually substitutes. For instance, the DM may believe that her predictions have better sensitivity but worse specificity than those of the machine, while 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 she is most certain a priori about whether an action is required for a task (i.e., when prior p takes more extreme values with $p < p^{\mathsf{B}}$ or $p > p^{\mathsf{W}}$). Conversely, the DM always properly learns the machine's type when she is most uncertain about whether or not act (i.e., prior p takes moderate values), as stated by the next corollary of Theorems 1, 2 and 3.

COROLLARY 2. *The DM always correctly learns the true type of the machine if and only if* $p \in (p^B, p^W)$*, whether the DM can override the machine or not.*

6.2. Learning with Anticipation

In our set-up, as in the literature, the DM updates her 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 3a), and always converges when it is worse (see Figure 4). Thus, the longer the DM remains uncertain (in the sense of Theorem 2), the more likely 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 she is correct in her assessment.

Assessing if the DM is correct when she 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 3b) or worse (see Figure 4b). To circumvent this issue, one approach consists in 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 2). However if the machine is worse, the convergence to either zero or one is random (per Theorem 3). 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), 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 will become increasingly confident about her 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 2-3). 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 her adoption decision over time. This is because the DM's belief in a better machine decreases in expectation over time and always approaches 0 in the limit in this case (per Theorems 2-3). 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 set-up with overriding oscillates and is recurrent in (0,1) (per Theorem 2 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 2-3 is that they do not stem form an inherent mistrust against the machine. Instead, they stem from four fundamentals (verification bias, exploration-free decisions, informativeness and substitution), which characterize the set-up, in which the DM works with the machine. Nonetheless, the DM may also be subject to mistrust biases against the machine in situations where these fundamentals are at play. In this section, we explore to which extent biases such as these interact with our four fundamentals to affect the DM's learning behavior.

In our main set-up, 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 she 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, Boyaci et al. 2020). 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

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

where $\lambda \in (0,1)$ represents the DM's mistrust bias against the machine's signal. The higher the value of λ is, 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 $\mathbb{P}(\Theta = \mathsf{A} \,|\, S^{\mathsf{H}}, S^{\mathsf{M}}, b_{t-1})$ derived from Bayes' rule in the main set-up. 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 set-up, the DM always overrules a better machine (never overrules a worse machine) if the bias is too high (resp., 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, where $\tilde{\mathbb{P}}^{\mathsf{B}}_\lambda(s^{\mathsf{H}}, s^{\mathsf{M}}) \triangleq \tilde{\mathbb{P}}_\lambda(s^{\mathsf{H}}, s^{\mathsf{M}}, 1)$ 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 λ_{min} and λ_{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 4. *Assume* $\lambda \in (\lambda_{min}, \lambda_{max})$.

- *When* $\Gamma = \mathsf{B}$*, if* $p \leq p^{\mathsf{B}}$*, then* b_t *oscillates and is recurrent; otherwise,* $b_t \xrightarrow{a.s.} 1$.
- *When* $\Gamma = \mathsf{W}$*, if* $p \leq p^{\mathsf{W}}$*, 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 2-3 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 appendix).

7.2. Mistrusting the Machine When Updating Belief

The DM's mistrust against the machine can also affect the way the DM updates her 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 her belief following Bayes' rule when the machine is correct but magnifies the decrease in belief when the machine is wrong. More formally, we have

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

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 set-up 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 5 **(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

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

Note that thresholds μ^{B} and μ^{W} actually play the same role as p^{B} and p^{W} in Theorems 2 and 3, respectively. Indeed, thresholds p^{B}_μ and p^{W}_μ exist such that $\mu > \mu^\Gamma \Leftrightarrow p < p^\Gamma_\mu$, for $\Gamma = \{\mathsf{B}, \mathsf{W}\}$. In particular, the structure of the results when the machine is worse (see Theorem 3) 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^{W}_μ.

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 3). This is in contrast to the main set-up 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 set-up 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 5 reduces to Theorems 2-3 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—while the DM always properly learns that the machine is better when she is not biased against the machine.

8. Conclusion

This paper proposes a framework in which a machine performs repeated decision tasks under the supervision of a DM. In this set-up, we fully characterize the evolution of the DM's belief about the

machine and overruling decisions over time. These results shed light on fundamental limitations in our ability to learn whether a machine, once deployed, makes better predictions than a human expert. Our findings further characterize the different forms that mislearning the true nature of the machine can take, which we trace back to the DM's overriding decisions. This provides a novel explanation for the joint production of decisions by machines and experts, and suggests several guidelines for properly learning whether to overrule the machine.

The learning failures we characterize in this paper do not arise from an intrinsic mistrust bias against machine-based predictions, such as the algorithmic aversion. Rather, they stem from the problem of learning about a machine while actually using its predictions to make high-stake decisions. We capture the key features of this situation 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 DM fully observes, or can act for the purpose of observing the machine's type. In contrast, we expect mislearning to occur even when some of the signals are not informative, or some complementarity exists between the machine and the DM. Note, however, that the problem may then become degenerated (when none of the signal are informative, for instance). In any case, replicating our analysis to these situations should be possible, and similar results should hold.

We also restrict our analysis to two possible machine's types, mostly for simplicity. But our framework can be extended to account for more types. Our results should not change overall as long as the previous fundamentals hold. Indeed, the DM's belief that the machine outperforms her expertise is what fundamentally matters when deciding to override the machine. This, in essence, divides the different possible machine's types into two distinct partitions depending on whether or not 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 set-up. Our results indicate that these biases can interact with our results in a significant way. In particular, the presence of a 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, to cite a few (Benjamin 2019). Further, our framework can potentially account for situations in which the DM does not perfectly know her own accuracy, or have 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's 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 where 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. While 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. 2021 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.

References

Acemoglu, D., Dahleh, M. A., Lobel, I. and Ozdaglar, A. (2011), 'Bayesian learning in social networks', *The Review of Economic Studies* **78**(4), 1201–1236.

Ahsen, M. E., Ayvaci, M. U. S. and Raghunathan, S. (2019), 'When algorithmic predictions use human-generated data: A bias-aware classification algorithm for breast cancer diagnosis', *Information Systems Research* **30**(1), 97–116.

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

Bastani, H., Bastani, O. and Sinchaisri, W. P. (2021), 'Learning best practices: Can machine learning improve human decision-making?'.

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

Baumeister, R. F., Bratslavsky, E., Finkenauer, C. and Vohs, K. D. (2001), 'Bad is stronger than good', *Review of general psychology* **5**(4), 323–370.

Begg, C. B. and Greenes, R. A. (1983), 'Assessment of diagnostic tests when disease verification is subject to selection bias', *Biometrics* pp. 207–215.

Benjamin, D. J. (2019), 'Errors in probabilistic reasoning and judgment biases', *Handbook of Behavioral Economics: Applications and Foundations 1* **2**, 69–186.

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

Boyaci, T., Canyakmaz, C. and de Véricourt, F. (2020), 'Human and machine: The impact of machine input on decision-making under cognitive limitations'.

Boyacı, T. and Özer, Ö. (2010), 'Information acquisition for capacity planning via pricing and advance selling: When to stop and act?', *Operations Research* **58**(5), 1328–1349.

Broemeling, L. D. (2011), 'Bayesian estimation of combined accuracy for tests with verification bias', *Diagnostics* **1**(1), 53–76.

Cheung, W. C., Simchi-Levi, D. and Wang, H. (2017), 'Dynamic pricing and demand learning with limited price experimentation', *Operations Research* **65**(6), 1722–1731.

Chung, K. L. and Zhong, K. (2001), *A course in probability theory*, Academic press.

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

Cowgill, B. (2019), 'Bias and productivity in humans and machines', *Columbia Business School Research Paper Forthcoming* .

Cukier, K., Mayer-Schönberger, V. and de Véricourt, F. (2021), *Framers: Human advantage in an age of technology and turmoil*, Penguin.

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

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

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

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E. and 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* **20**.

Guo, Y., Zhang, C. and Yang, X. J. (2020), Modeling trust dynamics in human-robot teaming: A bayesian inference approach, *in* 'Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems', pp. 1–7.

Gut, A. (2009), *Stopped random walks*, Springer.

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

Herrera, H. and Hörner, J. (2013), 'Biased social learning', *Games and Economic Behavior* **80**, 131–146.

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

Kemperman, J. (1974), 'The oscillating random walk', *Stochastic Processes and their applications* **2**(1), 1–29.

Keskin, N. B. and Birge, J. R. (2019), 'Dynamic selling mechanisms for product differentiation and learning', *Operations research* **67**(4), 1069–1089.

Kubat, M. (2017), *An introduction to machine learning*, Vol. 2, Springer.

Lebovitz, S., Levina, N. and Lifshitz-Assaf, H. (2021), 'Is ai ground truth really "true"? the dangers of training and evaluating ai tools based on experts' know-what', *Management Information Systems Quarterly* .

Lebovitz, S., Lifshitz-Assaf, H. and 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', *Organization Science* .

McKendrick, J. (2021), 'Ai adoption skyrocketed over the last 18 months', https://hbr.org/2021/09/ai-adoption-skyrocketed-over-the-last-18-months [Online; accessed 18-February-2022].

Möbius, M. M., Niederle, M., Niehaus, P. and Rosenblat, T. S. (2022), 'Managing self-confidence: Theory and experimental evidence', *Management Science* p. Forthcoming.

Özer, Ö. and Zheng, Y. (2018), 'Trust and trustworthiness', *The Handbook of Behavioral Operations* pp. 489–523.

Pepe, M. S. (2003), *The statistical evaluation of medical tests for classification and prediction*, Oxford University Press, USA.

Puranam, P. and Tsetlin, I. (2021), 'Explainability as an optimal stopping problem: Implications for human-ai interaction'.

Ransohoff, D. F. and Feinstein, A. R. (1978), 'Problems of spectrum and bias in evaluating the efficacy of diagnostic tests', *New England Journal of Medicine* **299**(17), 926–930.

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

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

Soll, J. B. and Mannes, A. E. (2011), 'Judgmental aggregation strategies depend on whether the self is involved', *International Journal of Forecasting* **27**(1), 81–102.

Stone, M. (1961), 'The opinion pool', *The Annals of Mathematical Statistics* pp. 1339–1342.

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

Van Donselaar, K. H., Gaur, V., Van Woensel, T., Broekmeulen, R. A. and Fransoo, J. C. (2010), 'Ordering behavior in retail stores and implications for automated replenishment', *Management Sci.* **56**(5), 766–784.

Vatutin, V. A. and Wachtel, V. (2009), 'Local probabilities for random walks conditioned to stay positive', *Probability Theory and Related Fields* **143**(1), 177–217.

Wang, C., Zhang, C. and Yang, X. J. (2018), Automation reliability and trust: A bayesian inference approach, *in* 'Proceedings of the Human Factors and Ergonomics Society Annual Meeting', Vol. 62, SAGE Publications Sage CA: Los Angeles, CA, pp. 202–206.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M. et al. (2019), 'Do no harm: a roadmap for responsible machine learning for health care', *Nature medicine* **25**(9), 1337–1340.

Zhou, X.-h. (1993), 'Maximum likelihood estimators of sensitivity and specificity corrected for verification bias', *Communications in Statistics-Theory and Methods,* **22**(11), 3177–3198.

Appendix A: No-overriding Benchmark

Proof of Theorem 1. We prove this result by first deriving a recursive expression (in terms of b_{t-1}) for belief b_t using Bayes' rule. Then, we focus on the log-likelihood ratio process L_t defined by $L_t = \log(b_t/(1-b_t))$. Observe that when $L_t \to \infty$ (and/or $L_t \to -\infty$) almost surely, then it immediately follows that $b_t \to 1$ (and resp. $b_t \to 0$) due to the continuous mapping theorem since L_t is a continuous monotone transformation of b_t.

In the no-overriding benchmark, the DM acts only if $S_t^{\mathsf{M}} = +$, so it follows that

$$b_t = \frac{b_{t-1}\left(\alpha^{\mathsf{B}}\right)^{1_{\{S_t^{\mathsf{M}}=+,\Theta_t=\mathsf{A}\}}}\left(\bar{\beta}^{\mathsf{B}}\right)^{1_{\{S_t^{\mathsf{M}}=+,\Theta_t=\mathsf{NA}\}}}}{b_{t-1}\left(\alpha^{\mathsf{B}}\right)^{1_{\{S_t^{\mathsf{M}}=+,\Theta_t=\mathsf{A}\}}}\left(\bar{\beta}^{\mathsf{B}}\right)^{1_{\{S_t^{\mathsf{M}}=+,\Theta_t=\mathsf{NA}\}}} + \bar{b}_{t-1}\left(\alpha^{\mathsf{W}}\right)^{1_{\{S_t^{\mathsf{M}}=+,\Theta_t=\mathsf{A}\}}}\left(\bar{\beta}^{\mathsf{W}}\right)^{1_{\{S_t^{\mathsf{M}}=+,\Theta_t=\mathsf{NA}\}}}} \tag{13}$$

where $1_{\{\}}$ is the indicator variable. Using the definition of L_t, we obtain $L_t = L_{t-1} + R_t^{\mathsf{M}}$, where

$$R_t^{\mathsf{M}} \triangleq 1_{\{S_t^{\mathsf{M}}=+,\Theta_t=\mathsf{A}\}} \log\left(\frac{\alpha^{\mathsf{B}}}{\alpha^{\mathsf{W}}}\right) + 1_{\{S_t^{\mathsf{M}}=+,\Theta_t=\mathsf{NA}\}} \log\left(\frac{\bar{\beta}^{\mathsf{B}}}{\bar{\beta}^{\mathsf{W}}}\right). \tag{14}$$

Therefore, L_t is a random walk with i.i.d. random jumps R_t^{M}.

When the machine's type is $\Gamma \in \{\mathsf{B}, \mathsf{W}\}$, then the mean of the random jump is

$$\mathbb{E}^{\Gamma}[R_t^{\mathsf{M}}] = p\alpha^{\Gamma} \log\left(\frac{\alpha^{\mathsf{B}}}{\alpha^{\mathsf{W}}}\right) + \bar{p}\bar{\beta}^{\Gamma} \log\left(\frac{\bar{\beta}^{\mathsf{B}}}{\bar{\beta}^{\mathsf{W}}}\right). \tag{15}$$

If $p < p^{\Gamma}$, it follows that the mean $\mathbb{E}^{\Gamma}[R_t^{\mathsf{M}}]$ and, hence, the drift of the random walk L_t is negative so $L_t \to -\infty$ (see Gut 2009, Theorem 9.1). The reverse condition (with strict inequality) implies the divergence to ∞. If the mean $\mathbb{E}^{\Gamma}[R_t^{\mathsf{M}}]$ equals 0 ($p = p^{\Gamma}$), then L_t is a martingale; hence, b_t oscillates (see Theorem 8.3.4 in Chung and Zhong 2001).

Finally, p^{B} and p^{W} are such that $p^{\mathsf{B}} < p^{\mathsf{W}}$ because $\alpha^{\mathsf{B}}/\bar{\beta}^{\mathsf{B}} > \alpha^{\mathsf{W}}/\bar{\beta}^{\mathsf{W}}$, which is implied by Substitution (4)-(5). Q.E.D.

Appendix B: Learning with Overriding

Proof of Lemma 1. This lemma follows from the fact that posterior probabilities are continuous and monotone in b_{t-1} and the boundary values 1 and 0 are on different sides of r due to Substitution (4)-(5). In particular, the posterior probabilities are

$$\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^{\mathsf{H}} = +, S_t^{\mathsf{M}} = -, b_{t-1}) = \frac{\alpha^{\mathsf{H}}(b_{t-1}\bar{\alpha}^{\mathsf{B}} + \bar{b}_{t-1}\bar{\alpha}^{\mathsf{W}})p}{\alpha^{\mathsf{H}}(b_{t-1}\bar{\alpha}^{\mathsf{B}} + \bar{b}_{t-1}\bar{\alpha}^{\mathsf{W}})p + \bar{\beta}^{\mathsf{H}}(b_{t-1}\beta^{\mathsf{B}} + \bar{b}_{t-1}\beta^{\mathsf{W}})\bar{p}},$$
$$\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^{\mathsf{H}} = -, S_t^{\mathsf{M}} = +, b_{t-1}) = \frac{\bar{\alpha}^{\mathsf{H}}(b_{t-1}\alpha^{\mathsf{B}} + \bar{b}_{t-1}\alpha^{\mathsf{W}})p}{\bar{\alpha}^{\mathsf{H}}(b_{t-1}\alpha^{\mathsf{B}} + \bar{b}_{t-1}\alpha^{\mathsf{W}})p + \beta^{\mathsf{H}}(b_{t-1}\bar{\beta}^{\mathsf{B}} + \bar{b}_{t-1}\bar{\beta}^{\mathsf{W}})\bar{p}}.$$

By solving the following equations for b_{t-1},

$$\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^{\mathsf{H}} = +, S_t^{\mathsf{M}} = -, b_{t-1}) = r \text{ and } \mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^{\mathsf{H}} = -, S_t^{\mathsf{M}} = +, b_{t-1}) = r \tag{16}$$

we obtain the thresholds

$$b^- = \frac{\left(\frac{\bar{r}p\alpha^{\mathsf{H}}}{r\bar{p}\bar{\beta}^{\mathsf{H}}}\right)\bar{\alpha}^{\mathsf{W}} - \beta^{\mathsf{W}}}{\left(\frac{\bar{r}p\alpha^{\mathsf{H}}}{r\bar{p}\bar{\beta}^{\mathsf{H}}}\right)\left[\alpha^{\mathsf{B}} - \alpha^{\mathsf{W}}\right] + \beta^{\mathsf{B}} - \beta^{\mathsf{W}}} \text{ and } b^+ = \frac{\bar{\beta}^{\mathsf{W}} - \left(\frac{\bar{r}p\bar{\alpha}^{\mathsf{H}}}{r\bar{p}\beta^{\mathsf{H}}}\right)\alpha^{\mathsf{W}}}{\left(\frac{\bar{r}p\bar{\alpha}^{\mathsf{H}}}{r\bar{p}\beta^{\mathsf{H}}}\right)\left[\alpha^{\mathsf{B}} - \alpha^{\mathsf{W}}\right] + \beta^{\mathsf{B}} - \beta^{\mathsf{W}}}. \tag{17}$$

Thus, the result follows. Q.E.D.

Before proving Theorems 2 and 3, we first provide two constructive lemmas and their proofs.

LEMMA 3. *With overriding, the log-likelihood ratio process L_t is as follows.*

Case 1: If $b^+ > b^-$, then we have $L_t = L_{t-1} + R_t^{HM}(L_{t-1})$, where

$$R_t^{HM}(L_{t-1}) \triangleq \begin{cases} 1_{\{S_t^M=+\}}\left[1_{\{\Theta_t=A\}}\log\left(\frac{\alpha^B}{\alpha^W}\right) + 1_{\{\Theta_t=NA\}}\log\left(\frac{\bar{\beta}^B}{\bar{\beta}^W}\right)\right] & \text{if } L_{t-1} \geq L_h \\ 1_{\{S_t^H=+,S_t^M=+\}}\left[1_{\{\Theta_t=A\}}\log\left(\frac{\alpha^B}{\alpha^W}\right) + 1_{\{\Theta_t=NA\}}\log\left(\frac{\bar{\beta}^B}{\bar{\beta}^W}\right)\right] & \text{if } L_h > L_{t-1} > L_\ell \\ 1_{\{S_t^H=+\}}\left[\upsilon_1 + \upsilon_2\right] & \text{if } L_\ell \geq L_{t-1} \end{cases}$$

with $L_h \triangleq \log\left(\frac{b^+}{\bar{b}^+}\right)$, $L_\ell \triangleq \log\left(\frac{b^-}{\bar{b}^-}\right)$ and

$$\upsilon_1 \triangleq 1_{\{\Theta_t=A\}}\left(1_{\{S_t^M=-\}}\log\left(\frac{\bar{\alpha}^B}{\bar{\alpha}^W}\right) + 1_{\{S_t^M=+\}}\log\left(\frac{\alpha^B}{\alpha^W}\right)\right),$$

$$\upsilon_2 \triangleq 1_{\{\Theta_t=NA\}}\left(1_{\{S_t^M=-\}}\log\left(\frac{\beta^B}{\beta^W}\right) + 1_{\{S_t^M=+\}}\log\left(\frac{\bar{\beta}^B}{\bar{\beta}^W}\right)\right).$$

Case 2: If $b^+ \leq b^-$, then we have $L_t = L_{t-1} + R_t^{HM}(L_{t-1})$ where

$$R_t^{HM}(L_{t-1}) \triangleq \begin{cases} 1_{\{S_t^M=+\}}\left[1_{\{\Theta_t=A\}}\log\left(\frac{\alpha^B}{\alpha^W}\right) + 1_{\{\Theta_t=NA\}}\log\left(\frac{\bar{\beta}^B}{\bar{\beta}^W}\right)\right] & \text{if } L_{t-1} > L_h \\ 1_{\{\Theta_t=A\}}\nu_1 + 1_{\{\Theta_t=NA\}}\nu_2 & \text{if } L_h \geq L_{t-1} \geq L_\ell \\ 1_{\{S_t^H=+\}}\left[\upsilon_1 + \upsilon_2\right] & \text{if } L_\ell > L_{t-1} \end{cases} \tag{18}$$

with $L_h \triangleq \log\left(\frac{b^-}{\bar{b}^-}\right)$, $L_\ell \triangleq \log\left(\frac{b^+}{\bar{b}^+}\right)$ and

$$\nu_1 \triangleq 1_{\{S_t^M=+\}}\log\left(\frac{\alpha^B}{\alpha^W}\right) + 1_{\{S_t^M=-\}}1_{\{S_t^H=+\}}\log\left(\frac{\bar{\alpha}^B}{\bar{\alpha}^W}\right),$$

$$\nu_2 \triangleq 1_{\{S_t^M=+\}}\log\left(\frac{\bar{\beta}^B}{\bar{\beta}^W}\right) + 1_{\{S_t^M=-\}}1_{\{S_t^H=+\}}\log\left(\frac{\beta^B}{\beta^W}\right).$$

Proof of Lemma 3. As in the proof of Theorem 1, we derive a recursive expression for b_t in terms of b_{t-1} using Bayes' rule but this time using the decision-making procedure characterized in Lemma 1.

Case 1. When $b^+ > b^-$, we have the following three regimes

- $b_{t-1} \geq b^+$:

$$\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^{\mathsf{H}} = -, S_t^{\mathsf{M}} = +, b_{t-1}) \geq r \tag{19}$$

$$\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^{\mathsf{H}} = +, S_t^{\mathsf{M}} = -, b_{t-1}) < r \tag{20}$$

In this case, $S_t^{\mathsf{M}} = +$ is sufficient and necessary to act, which implies the machine overrides the human's signal $S_t^{\mathsf{H}} = -$. Further, $S_t^{\mathsf{H}} = +$ is also overruled by $S^{\mathsf{M}} = -$.

- $b^+ > b_{t-1} > b^-$:

$$\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^{\mathsf{H}} = -, S_t^{\mathsf{M}} = +, b_{t-1}) < r \tag{21}$$

$$\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^{\mathsf{H}} = +, S_t^{\mathsf{M}} = -, b_{t-1}) < r \tag{22}$$

In this case, $S_t^{\mathsf{M}} = S_t^{\mathsf{H}} = +$ is the only condition for acting.

- $b^- \geq b_{t-1}$:

$$\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^{\mathsf{H}} = -, S_t^{\mathsf{M}} = +, b_{t-1}) < r \tag{23}$$

$$\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^{\mathsf{H}} = +, S_t^{\mathsf{M}} = -, b_{t-1}) \geq r \tag{24}$$

In this case, $S_t^{\mathsf{H}} = +$ is sufficient and necessary to act. The signal of the human overrides the machine's signal in both conflicting cases.

Finally, the belief update when $b^+ > b^-$ is as follows.

$$b_t = \begin{cases} \left[1 + \frac{\bar{b}_{t-1}}{b_{t-1}} \left(\frac{\alpha^{\mathsf{W}}}{\alpha^{\mathsf{B}}}\right)^{1_{\{\Theta_t = \mathsf{A}, S_t^{\mathsf{M}} = +\}}} \left(\frac{\bar{\beta}^{\mathsf{W}}}{\bar{\beta}^{\mathsf{B}}}\right)^{1_{\{\Theta_t = \mathsf{NA}, S_t^{\mathsf{M}} = +\}}}\right]^{-1} & \text{if } b_{t-1} \geq b^+ \\ \left[1 + \frac{\bar{b}_{t-1}}{b_{t-1}} \left(\frac{\alpha^{\mathsf{W}}}{\alpha^{\mathsf{B}}}\right)^{1_{\{\Theta_t = \mathsf{A}, S_t^{\mathsf{M}} = +, S_t^{\mathsf{H}} = +\}}} \left(\frac{\bar{\beta}^{\mathsf{W}}}{\bar{\beta}^{\mathsf{B}}}\right)^{1_{\{\Theta_t = \mathsf{NA}, S_t^{\mathsf{M}} = +, S_t^{\mathsf{H}} = +\}}}\right]^{-1} & \text{if } b^+ > b_{t-1} > b^- \\ \left[1 + \frac{\bar{b}_{t-1}}{b_{t-1}} \zeta\right]^{-1} & \text{if } b^- \geq b_{t-1} \end{cases} \tag{25}$$

where

$$\zeta \triangleq \left\{ \left[\left(\frac{\bar{\alpha}^{\mathsf{W}}}{\bar{\alpha}^{\mathsf{B}}}\right)^{1_{\{S_t^{\mathsf{M}} = -\}}} \left(\frac{\alpha^{\mathsf{W}}}{\alpha^{\mathsf{B}}}\right)^{1_{\{S_t^{\mathsf{M}} = +\}}} \right]^{1_{\{\Theta_t = \mathsf{A}\}}} \left[\left(\frac{\beta^{\mathsf{W}}}{\beta^{\mathsf{B}}}\right)^{1_{\{S_t^{\mathsf{M}} = -\}}} \left(\frac{\bar{\beta}^{\mathsf{W}}}{\bar{\beta}^{\mathsf{B}}}\right)^{1_{\{S_t^{\mathsf{M}} = +\}}} \right]^{1_{\{\Theta_t = \mathsf{NA}\}}} \right\}^{1_{\{S_t^{\mathsf{H}} = +\}}} \tag{26}$$

Case 2. When $b^+ \leq b^-$ we have the following three regimes

- $b_{t-1} > b^-$:

$$\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^{\mathsf{H}} = -, S_t^{\mathsf{M}} = +, b_{t-1}) \geq r \tag{27}$$

$$\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^{\mathsf{H}} = +, S_t^{\mathsf{M}} = -, b_{t-1}) < r \tag{28}$$

In this case, $S_t^{\mathsf{M}} = +$ is sufficient and necessary to act, which implies the machine overrides the human's signal $S_t^{\mathsf{H}} = -$. Further, $S_t^{\mathsf{H}} = +$ is also overruled by $S_t^{\mathsf{M}} = -$.

- $b^- \geq b_{t-1} \geq b^+$:

$$\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^{\mathsf{H}} = -, S_t^{\mathsf{M}} = +, b_{t-1}) \geq r \tag{29}$$

$$\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^{\mathsf{H}} = +, S_t^{\mathsf{M}} = -, b_{t-1}) \geq r \tag{30}$$

In this case, $S_t^{\mathsf{M}} = +$ or $S_t^{\mathsf{H}} = +$ is sufficient for acting.

- $b^+ > b_{t-1}$:

$$\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^{\mathsf{H}} = -, S_t^{\mathsf{M}} = +, b_{t-1}) < r \tag{31}$$

$$\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^{\mathsf{H}} = +, S_t^{\mathsf{M}} = -, b_{t-1}) \geq r \tag{32}$$

In this case, $S_t^{\mathsf{H}} = +$ is sufficient and necessary to act. The signal of the human overrides the machine's signal in both conflicting cases.

Finally the belief update if $b^- \leq b^+$ is as follows.

$$b_t = \begin{cases} \left[1 + \frac{\bar{b}_{t-1}}{b_{t-1}} \left(\frac{\alpha^{\mathsf{W}}}{\alpha^{\mathsf{B}}}\right)^{1_{\{\Theta_t = \mathsf{A}, S_t^{\mathsf{M}} = +\}}} \left(\frac{\bar{\beta}^{\mathsf{W}}}{\bar{\beta}^{\mathsf{B}}}\right)^{1_{\{\Theta_t = \mathsf{NA}, S_t^{\mathsf{M}} = +\}}}\right]^{-1} & \text{if } b_{t-1} > b^- \\ \left[1 + \frac{\bar{b}_{t-1}}{b_{t-1}} \iota\right]^{-1} & \text{if } b^- \geq b_{t-1} \geq b^+ \\ \left[1 + \frac{\bar{b}_{t-1}}{b_{t-1}} \zeta\right]^{-1} & \text{if } b^+ > b_{t-1} \end{cases} \tag{33}$$

where ζ is defined in (26) and

$$\iota \triangleq \left[\left(\frac{\alpha^{\mathsf{W}}}{\alpha^{\mathsf{B}}}\right)^{1_{\{S_t^{\mathsf{M}} = +\}}} \left(\frac{\bar{\alpha}^{\mathsf{W}}}{\bar{\alpha}^{\mathsf{B}}}\right)^{1_{\{S_t^{\mathsf{M}} = -, S_t^{\mathsf{H}} = +\}}}\right]^{1_{\{\Theta_t = \mathsf{A}\}}} \left[\left(\frac{\bar{\beta}^{\mathsf{W}}}{\bar{\beta}^{\mathsf{B}}}\right)^{1_{\{S_t^{\mathsf{M}} = +\}}} \left(\frac{\beta^{\mathsf{W}}}{\beta^{\mathsf{B}}}\right)^{1_{\{S_t^{\mathsf{M}} = -, S_t^{\mathsf{H}} = +\}}}\right]^{1_{\{\Theta_t = \mathsf{NA}\}}}$$

The log-likelihood ratio process L_t is then obtained by $L_t = \log\left(b_t / \bar{b}_t\right)$ in both cases. Q.E.D.

LEMMA 4. *Consider the following sequences of random variables*

- *i.i.d.* $Y_{1,t}$ *with* $\mathbb{E}[Y_{1,t}] > 0$ *and* $|Y_{1,t}| \leq Y_{1,h}$ *where* $\infty > Y_{1,h} > 0$ *for* $t = 1, \ldots,$
- *i.i.d.* $Y_{2,t}$ *with* $\mathbb{P}(Y_{2,t} > 0) > 0$ *and* $|Y_{2,t}| \leq Y_{2,h}$ *where* $\infty > Y_{2,h} > 0$ *for* $t = 1, \ldots,$
- *and, i.i.d.* $Y_{3,t}$ *with* $\mathbb{E}[Y_{3,t}] \geq 0$ *and* $|Y_{3,t}| \leq Y_{3,h}$ *where* $\infty > Y_{3,h} > 0$ *for* $t = 1, \ldots.$

Let Z_t *be a discrete stochastic process governed by* $Y_{i,t}$ *and two thresholds* Z_ℓ *and* Z_h *such that* $Z_h > Z_\ell$ *as follows.*

$$Z_{t+1} = Z_t + Y_{1,t} 1_{\{Z_t \geq Z_h\}} + Y_{2,t} 1_{\{Z_t \in (Z_\ell, Z_h)\}} + Y_{3,t} 1_{\{Z_t \leq Z_\ell\}}. \tag{34}$$

Then, $Z_t \xrightarrow{a.s.} \infty$.

Proof of Lemma 4. Note that the divergence immediately follows when the mean of $Z_{t+1} - Z_t$ is positive for any Z_t because in that case in all three regions process Z_t is driven by random walks drifting to ∞. Thus, we focus on the most extreme parameter regime in terms of divergence to ∞, in particular, when the mean of $Z_{t+1} - Z_t$ is nonpositive in (Z_ℓ, Z_h) and 0 for $Z_t \leq Z_\ell$. To prove this result, we show that there exists a finite (but random) period τ such that process Z_t remains above $Z_t \geq Z_h$ for all $t > \tau$. This is sufficient for divergence to ∞ because process Z_t is governed

by a random walk drifting to ∞ above Z_h. We prove this in two steps, but first, we define the following stopping times recursively by assuming, without loss of generality, that $Z_0 > Z_h$.

$$T_1 = \inf\{t : Z_t < Z_h\}, \tag{35}$$

$$\tilde{T}_1 = \inf\{t > T_1 : Z_t \geq L_h\}, \tag{36}$$

$$T_i = \inf\{t > \tilde{T}_{i-1} : Z_t < Z_h\} \ \forall i > 1, \tag{37}$$

$$\tilde{T}_i = \inf\{t > T_i : Z_t \geq Z_h\} \ \forall i > 1\,. \tag{38}$$

Here, if one of the sets above is empty, i.e., no such t exists, we assign $T_i = \infty$ (and respectively $\tilde{T}_i = \infty$) for the corresponding set.

Step 1. In the first step, we prove that $\mathbb{P}(\tilde{T}_i < \infty \,|\, T_i < \infty) = 1$ for all i. In particular, if process Z_t goes below Z_h once, then with probability one, it will cross up Z_h in finite steps. This result also implies that the sequence of infinite stopping times (if it exists) is started by $T_j = \infty$ (but not $\tilde{T}_j = \infty$) for some j. To prove this result, first fix i and assume that T_i is finite; then we define a new sequence of stopping times for process Z_t.

$$V_1 = \inf\{t > T_i : Z_t \leq Z_\ell\} \tag{39}$$

$$\tilde{V}_1 = \inf\{t > V_1 : Z_t > Z_\ell\} \tag{40}$$

$$V_i = \inf\{t > \tilde{V}_{i-1} : Z_t \leq Z_\ell\} \ \forall i > 1, \tag{41}$$

$$\tilde{V}_i = \inf\{t > V_i : Z_t > Z_\ell\} \ \forall i > 1. \tag{42}$$

First, note that V_i and $\tilde{V}_i$ for $i \geq 1$ are proper random variables, i.e., they take finite values with probability one. This is because

$$\mathbb{P}(V_i < \infty \,|\, \tilde{V}_{i-1} < \infty) = 1 \tag{43}$$

$$\mathbb{P}(\tilde{V}_i < \infty \,|\, V_i < \infty) = 1\,. \tag{44}$$

The first equality holds because as discussed at the beginning of this proof, we focus on the case where the mean of $Z_{t+1} - Z_t$ in (Z_ℓ, Z_h) is either negative or 0. If negative, Z_t is governed by a random walk drifting to $-\infty$ in (Z_ℓ, Z_h); thus it crosses Z_ℓ with probability one in finite steps (see, Theorem 9.1 in Gut 2009, p. 70). If 0, Z_t in (Z_ℓ, Z_h) is a martingale and oscillates (see, Theorem 8.3 in Gut 2009, p. 68). The second equality holds because Z_t is governed again by an oscillating random walk when $Z_t < Z_\ell$ thus it crosses Z_ℓ in finite steps. Furthermore, $T_i < \infty$; thus, V_1 is also finite. Therefore, it follows that $V_i < \infty$ and $\tilde{V}_i < \infty$ for $i \geq 1$.

To prove that $\tilde{T}_i$ is finite, we will show that there exists a finite j such that $\tilde{T}_i < V_j$. To do so, define events $A_t = \{\tilde{T}_i \geq V_t\}$. Then, the Borel-Cantelli lemma implies the following,

$$\sum_{t=T_i}^{\infty} \mathbb{P}(A_t) < \infty \Rightarrow \mathbb{P}(\limsup_{i \to \infty} A_t) = 0 \tag{45}$$

First, consider $\mathbb{P}(A_1) = \mathbb{P}(\tilde{T}_i \geq V_1)$, this probability is bounded, i.e., $\mathbb{P}(A_1) < \delta < 1$ because $Z_h - Z_\ell$ is bounded and Z_t has positive size jumps with positive probability in (Z_ℓ, Z_h). Proceeding similarly, we obtain the following

$$\mathbb{P}(\tilde{T}_i \geq V_t) = \mathbb{P}(\tilde{T}_i \geq V_{t-1})\mathbb{P}(\tilde{T}_i \geq V_t \,|\, \tilde{T}_i \geq V_{t-1}) < \delta^t \tag{46}$$

Therefore, it follows that $\mathbb{P}(\limsup_{i\to\infty} A_t) = 0$, which implies that there exists a finite j such that $\tilde{T}_i < V_j$. As discussed, V_j is finite; thus, it follows that $\mathbb{P}(\tilde{T}_i < \infty \,|\, T_i < \infty) = 1$.

Step 2. In this step, we show that there exists a finite j such that $T_j = \infty$ and $\tilde{T}_k < \infty$ for all $k < j$. Define the following events $E_t = \{T_t < \infty\}$ for $t > 1$. Then, the Borel-Cantelli lemma implies the following

$$\sum_{t=1}^{\infty} \mathbb{P}(E_t) < \infty \Rightarrow \mathbb{P}(\limsup_{t\to\infty} E_t) = 0\,. \tag{47}$$

In particular, if the summation condition is satisfied, then events E_t cannot occur infinitely many times, i.e., there exists a finite j such that $T_j = \infty$. To show that the summation condition is indeed satisfied, we construct a finite upper bound for it. Above Z_h, process Z_t is driven by a random walk drifting to ∞. Thus, stopping time T_1 is defective, i.e., $\mathbb{P}(T_1 < \infty) < \varphi < 1$ for some φ (see Theorem 9.1 in Gut 2009, p.70). Moreover, we bound $\mathbb{P}(T_2 < \infty)$ as follows:

$$\mathbb{P}(T_2 < \infty) = \mathbb{P}(T_2 < \infty \,|\, \tilde{T}_1 < \infty)\mathbb{P}(\tilde{T}_1 < \infty \,|\, T_1 < \infty)\mathbb{P}(T_1 < \infty) < \varphi^2\,. \tag{48}$$

In Step 1 of this proof, we show that $\mathbb{P}(\tilde{T}_1 < \infty \,|\, T_1 < \infty) = 1$. Moreover, we have $\mathbb{P}(T_2 < \infty \,|\, \tilde{T}_1 < \infty) < \varphi < 1$ because $\tilde{T}_1 < \infty$ implies that process Z_t crosses Z_h up in finite steps, and after crossing Z_h, process Z_t is again driven by the same random walk drifting to ∞. Proceeding similarly, it follows that $\mathbb{P}(E_t) < \varphi^t$. Hence, there exists a finite (but random) j such that $T_j = \infty$, and Step 1 implies that $\tilde{T}_k < \infty$ for $k < j$ because $T_k < \infty$. Therefore, after sufficiently large t, process Z_t always remains above Z_h and diverges to ∞. Q.E.D.

Proof of Theorem 2. In this proof, we focus on the log-likelihood ratio process L_t because as also discussed in the proof of Theorem 1, the continuous mapping theorem implies that the limit of L_t characterizes the limit of b_t. Our proof is based on analyzing the mean of $L_{t+1} - L_t$ when L_t takes different values in comparison to L_h and L_ℓ for Case 1 and Case 2 characterized in Lemma 3. In both cases, process L_t is governed by different random walks with random jumps whose means and size change depending on the previous state L_{t-1}.[5] Trivial cases arise when the mean of $L_{t+1} - L_t$ always remains positive regardless of L_t. In particular, L_t diverges to ∞ when the mean of $L_{t+1} - L_t$ is always positive despite changing values of L_t due to the strong law of large

[5] The log-likelihood ratio process L_t is similar to the *oscillating random walk* defined in Kemperman (1974) such that the partition of $\mathbb{R}$ consists of $(-\infty, L_\ell]$, (L_ℓ, L_h) and $[L_h, \infty)$.

numbers because three random walks (for $L_t \geq L_h$, $L_t \in (L_h, L_\ell)$ and $L_t \leq L_\ell$ under Case 1) that establish the trajectory of process L_t all diverge to ∞ (see Theorem 8.3 in Gut 2009, p. 68). As a result, b_t converges to 0 as L_t diverges to ∞. We consider the different values of p in the statement of the theorem separately.

Step 1. Let $p \leq p^{\mathsf{B}}$. To prove that L_t and hence b_t oscillate, we need to show that there exists a number that process L_t crosses infinitely often (see, for instance, Vatutin and Wachtel 2009 for a mathematical definition of oscillation). This property (oscillation) holds for a random walk with a noise term whose mean is 0 (see Theorem 8.2 in Gut 2009, p. 68). However, process L_t does not satisfy this property because the mean of $L_{t+1} - L_t$ is always positive when $L_t \leq L_\ell$ in Case 1 ($L_t < L_\ell$ in Case 2) as discussed. Nevertheless, we can show that process L_t oscillates because the mean of $L_{t+1} - L_t$ is either negative or zero when $p \leq p^{\mathsf{B}}$ for $L_t \geq L_h$ in Case 1 ($L_t > L_h$ in Case 2). Therefore, process L_t returns to interval (L_ℓ, L_h) in finite steps after lying outside it. Oscillation is regardless of the sign of the mean of $L_{t+1} - L_t$ in (L_ℓ, L_h) because process L_t goes out of (L_ℓ, L_h) in finite steps. Specifically, in finite steps, process L_t i) crosses L_h up if the mean of $L_{t+1} - L_t$ in (L_ℓ, L_h) is positive, ii) crosses L_ℓ down if it is negative, and iii) goes out of (L_ℓ, L_h) if it is zero.

To illustrate this process, we focus on Case 1 and the setting where the mean of $L_{t+1} - L_t$ in (L_ℓ, L_h) is negative. Define the following stopping times

$$J = \inf\{t \geq 1 : L_t \leq L_\ell\} \tag{49}$$

$$\tilde{J} = \inf\{t \geq 1 : L_t > L_\ell\} \tag{50}$$

If J and $\tilde{J}$ are proper random variables (i.e., $\mathbb{P}(J < \infty) = \mathbb{P}(\tilde{J} < \infty) = 1$), then process L_t oscillates. Assume $L_0 > L_\ell$ without loss of generality; then J is a proper (almost surely finite) random variable because process L_t is driven by a random walk drifting to $-\infty$ for $L_t > L_\ell$. Thus, in finite steps it will cross L_ℓ down. After J steps, process L_t is driven by a random walk drifting to ∞,[6] thus it crosses L_ℓ up in finite steps, so $\tilde{J} - J$ is a proper random variable. Since J is also proper, $\tilde{J} = \tilde{J} - J + J$ is also proper. Since L_t crosses L_ℓ infinitely often, b_t crosses $1/(1+\exp(L_\ell))$ infinitely often and, hence, oscillates. The same approach can be repeated for the remaining setting where the mean of $L_{t+1} - L_t$ is positive or zero in (L_ℓ, L_h).

The oscillation property also implies that the process L_t and hence b_t is recurrent because the mean of the random jumps is bounded. With probability 1, process L_t will revisit interval (L_ℓ, L_h) for Case 1 in finite steps interval $[L_\ell, L_h]$ for Case 2 in finite steps. Furthermore, intervals that can be reached from (L_ℓ, L_h) with positive probability are also recurrent, which implies Corollary 1.

[6] This claim can be proved by invoking Lemma A.2 in Harrison et al. (2012) to show that $\mathbb{E}^{\mathsf{B}}[R_t^{\mathsf{HM}}(L_{t-1})] > 0$ for $L_{t-1} \leq L_\ell$ in Case 1, and $L_{t-1} < L_\ell$ in Case 2.

Step 2. Let $p > p^{\mathsf{B}}$. Then, the mean of $L_{t+1} - L_t$ when $L_t \le L_\ell$ for Case 1 and $L_t < L_\ell$ for Case 2 is positive (see Footnote 6). Further, we know from the proof of Theorem 1 that the mean of $L_{t+1} - L_t$ when $L_t \ge L_h$ for Case 1 and $L_t > L_h$ for $p > p^{\mathsf{B}}$ is also positive. In Case 1, the mean of $L_{t+1} - L_t$ for $L_t \in (L_\ell, L_h)$ is

$$\alpha^{\mathsf{H}}\alpha^{\mathsf{B}} p \log\left[\frac{\alpha^{\mathsf{B}}}{\alpha^{\mathsf{W}}}\right] + \bar{\beta}^{\mathsf{H}}\bar{\beta}^{\mathsf{B}}\bar{p}\log\left[\frac{\bar{\beta}^{\mathsf{B}}}{\bar{\beta}^{\mathsf{W}}}\right] = \bar{\beta}^{\mathsf{H}}\left[\alpha^{\mathsf{B}} p \log\left[\frac{\alpha^{\mathsf{B}}}{\alpha^{\mathsf{W}}}\right] + \bar{\beta}^{\mathsf{B}}\bar{p}\log\left[\frac{\bar{\beta}^{\mathsf{B}}}{\bar{\beta}^{\mathsf{W}}}\right]\right] + \bar{\beta}^{\mathsf{H}}\alpha^{\mathsf{B}}\left[\frac{\alpha^{\mathsf{H}}}{\bar{\beta}^{\mathsf{H}}} - 1\right]\log\left[\frac{\alpha^{\mathsf{B}}}{\alpha^{\mathsf{W}}}\right].$$

Note that the term above is positive for $p > p^{\mathsf{B}}$. Hence, as discussed L_t diverges to ∞ and b_t converges to 1 for Case 1 when p is larger than p^{B}. In Case 2, the mean of $L_{t+1} - L_t$ for $L_t \in [L_\ell, L_h]$ is not necessarily positive. When it is positive, we immediately obtain the same result. Nevertheless, we obtain the same divergence despite having negative or zero mean of $L_{t+1} - L_t$ for $L_t \in [L_\ell, L_h]$ as long as the means of $L_{t+1} - L_t$ for $L_t < L_\ell$ and $L_t > L_h$ are positive, which is true as proven in Lemma 4. Thus, using Lemma 4, we capture all possible values of L_t with respect to L_h and L_ℓ in Cases 1 and 2 for $p > p^{\mathsf{B}}$, and show that L_t diverges to ∞, which implies b_t converges to 1. Hence, we conclude the proof. Q.E.D.

Proof of Theorem 3. The first part of the result in this theorem when $p \le p^{\mathsf{W}}$ follows from Lemma 4 by considering the reflection of the stochastic process. In particular, updating the conditions in the statement of Lemma 4 as $\mathbb{E}[Y_{1,t}] \le 0$, $\mathbb{P}(Y_{2,t} < 0) > 0$ and $\mathbb{E}[Y_{3,t}] < 0$ would imply $Z_t \xrightarrow{\text{a.s.}} -\infty$ in that lemma. This is because the mean of $L_{t+1} - L_t$ is nonpositive when $L_t \in [L_h, \infty)$ in Case 1 (and $L_t \in (L_h, \infty)$ in Case 2); and is negative when $L_t \in (-\infty, L_\ell]$ in Case 1 (and $L_t \in (-\infty, L_\ell)$ in Case 2).[7] Thus, the case for $p \le p^{\mathsf{W}}$ follows from Lemma 4 with a slight modification.

To prove the second part of this result $p > p^{\mathsf{W}}$, we focus on Case 1 (Case 2 can be addressed by adjusting weak and strict inequalities in the same way) by assuming the mean of $L_{t+1} - L_t$ is negative when $L_t \in (L_\ell, L_h)$ and define the following stopping times by assuming $L_0 > L_h$ without loss of generality.

$$T_1 = \inf\{t : L_t < L_h\} \tag{51}$$

$$\tilde{T}_1 = \inf\{t > T_1 : L_t \ge L_h\} \tag{52}$$

$$T_i = \inf\{t > \tilde{T}_{i-1} : L_t < L_h\} \qquad \text{for } i \ge 2 \tag{53}$$

$$\tilde{T}_i = \inf\{t > T_i : L_t \ge L_h\} \qquad \text{for } i \ge 2 \tag{54}$$

Here, if a set is empty, then the stopping times takes the value of ∞. Therefore, if one of the stopping times T_i or $\tilde{T}_i$ is not finite for some i, then all the following stopping times for $j > i$ also are ∞. We first show that $\sum_{i=1}^{\infty} \mathbb{P}(T_i < \infty) < \infty$ and then use the Borel-Cantelli lemma to deduce

[7] The second part of this claim can be proved by invoking Lemma A.2 in Harrison et al. (2012) to show that $\mathbb{E}^{\mathsf{B}}[R_t^{\mathsf{HM}}(L_{t-1})] < 0$ for $L_{t-1} \le L_\ell$ in Case 1, and $L_{t-1} < L_\ell$ in Case 2.

that there exists a finite j such that $\mathbb{P}(T_j = \infty \text{ or } \tilde{T}_j = \infty) = 1$. If $T_j = \infty$, then $L_t \to \infty$; otherwise, $(\tilde{T}_j = \infty)$ then $L_t \to -\infty$. Thus, $L_t \to L$, where $L \in \{-\infty, \infty\}$; hence b_t converges to a Bernoulli random variable.

First, consider T_1; it follows that $\mathbb{P}(T_1 < \infty) < \xi < 1$ (i.e., T_1 is a defective random variable) because process L_t is driven by a random walk drifting to ∞ when $L_t > L_h$ and $p > p^{\mathsf{W}}$. Next, considering $\tilde{T}_1$ and T_2; we obtain that

$$\mathbb{P}(\tilde{T}_1 < \infty) = \mathbb{P}(\tilde{T}_1 < \infty \,|\, T_1 < \infty)\mathbb{P}(T_1 < \infty) < \xi^2 \tag{55}$$

$$\mathbb{P}(T_2 < \infty) = \mathbb{P}(T_2 < \infty \,|\, \tilde{T}_1 < \infty)\mathbb{P}(\tilde{T}_1 < \infty) < \xi^3 \tag{56}$$

Here, the first term is strictly less than 1, i.e., $\mathbb{P}(\tilde{T}_1 < \infty \,|\, T_1 < \infty) < \xi < 1$ because process L_t is driven by random walks drifting to $-\infty$ after T_1. The inequality in the second line follows because L_t is driven by a random walk drifting to ∞. Proceeding similarly, it follows that $\mathbb{P}(T_i < \infty) < \xi^{(2i-1)}$ and hence $\sum_{i=1}^{\infty} \mathbb{P}(T_i < \infty)$ is finite.

Note that we assume at the beginning that the mean of $L_{t+1} - L_t$ is negative when $L_t \in (L_\ell, L_h)$. If it is positive, the same steps can be followed by redefining stopping times using L_ℓ as the threshold instead of L_h. If it is zero, then stopping times T_i is defined by considering the time when process L_t enters (L_ℓ, L_h) and $\tilde{T}_i$ is the time when L_t exists (L_ℓ, L_h) in any direction. The approach follows because outside (L_ℓ, L_h), process L_t diverges (either to ∞ or $-\infty$ depending on being above L_h or below L_ℓ) and it oscillates in (L_ℓ, L_h), which guarantees it remains in (L_ℓ, L_h) for finite steps.

More specifically, the main driver of this result is that process L_t may diverge to ∞ when above L_h and to $-\infty$ when below L_ℓ. The sign of the mean between L_ℓ and L_h does not affect this characteristic in the limit. Thus, we conclude the proof. Q.E.D.

Appendix C: Mistrust Bias against the Machine

Proof of Lemma 2. Note that Informativeness (1)-(2) imply that the posterior after + (after −) is larger than or equal to r (resp. lower than r). Thus, there exist thresholds $\lambda_\Gamma^-, \lambda_\Gamma^+ \in (0,1)$ for $\Gamma \in \{\mathsf{B}, \mathsf{W}\}$ given by

$$\lambda_\Gamma^- \triangleq \frac{r - \mathbb{P}^\Gamma(\Theta_t = \mathsf{A} \,|\, S^{\mathsf{M}} = -)}{\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S^{\mathsf{H}} = +) - \mathbb{P}^\Gamma(\Theta_t = \mathsf{A} \,|\, S^{\mathsf{M}} = -)}, \tag{57}$$

$$\lambda_\Gamma^+ \triangleq \frac{\mathbb{P}^\Gamma(\Theta_t = \mathsf{A} \,|\, S^{\mathsf{M}} = +) - r}{\mathbb{P}^\Gamma(\Theta_t = \mathsf{A} \,|\, S^{\mathsf{M}} = +) - \mathbb{P}(\Theta_t = \mathsf{A} \,|\, S^{\mathsf{H}} = -)}. \tag{58}$$

These equations imply that

$$\lambda \mathbb{P}(\Theta_t = \mathsf{A} \,|\, S^{\mathsf{H}} = +) + \bar{\lambda}\mathbb{P}^{\mathsf{B}}(\Theta_t = \mathsf{A} \,|\, S^{\mathsf{M}} = -) \geq r \;\; \Leftrightarrow \lambda \geq \lambda_{\mathsf{B}}^-, \tag{59}$$

$$\lambda \mathbb{P}(\Theta_t = \mathsf{A} \,|\, S^{\mathsf{H}} = -) + \bar{\lambda}\mathbb{P}^{\mathsf{B}}(\Theta_t = \mathsf{A} \,|\, S^{\mathsf{M}} = +) \leq r \;\; \Leftrightarrow \lambda \geq \lambda_{\mathsf{B}}^+, \tag{60}$$

$$\lambda \mathbb{P}(\Theta_t = \mathsf{A} \,|\, S^{\mathsf{H}} = +) + \bar{\lambda}\mathbb{P}^{\mathsf{W}}(\Theta_t = \mathsf{A} \,|\, S^{\mathsf{M}} = -) \leq r \;\; \Leftrightarrow \lambda \leq \lambda_{\mathsf{W}}^-, \tag{61}$$

$$\lambda \mathbb{P}(\Theta_t = \mathsf{A} \,|\, S^{\mathsf{H}} = -) + \bar{\lambda}\mathbb{P}^{\mathsf{W}}(\Theta_t = \mathsf{A} \,|\, S^{\mathsf{M}} = +) \geq r \;\; \Leftrightarrow \lambda \leq \lambda_{\mathsf{W}}^+, \tag{62}$$

because the left-hand sides of the first and third inequalities are increasing and the second and the fourth ones are decreasing in λ.

First, note that λ_Γ^+ is increasing in $\mathbb{P}^\Gamma(\Theta_t = \mathsf{A} \,|\, S^\mathsf{M} = +)$, and λ_Γ^- is decreasing in $\mathbb{P}^\Gamma(\Theta_t = \mathsf{A} \,|\, S^\mathsf{M} = -)$. We also know from Substitution (4)-(5) that $\mathbb{P}^\mathsf{B}(\Theta_t = \mathsf{A} \,|\, S^\mathsf{M} = +) > \mathbb{P}^\mathsf{W}(\Theta_t = \mathsf{A} \,|\, S^\mathsf{M} = +)$, and $\mathbb{P}^\mathsf{W}(\Theta_t = \mathsf{A} \,|\, S^\mathsf{M} = -) > \mathbb{P}^\mathsf{B}(\Theta_t = \mathsf{A} \,|\, S^\mathsf{M} = -)$. Thus, the ranking between thresholds at the statement of the proposition follows. We conclude the proof by defining $\lambda_{min} \triangleq \max(\lambda_\mathsf{W}^-, \lambda_\mathsf{W}^+)$ and $\lambda_{max} = \min(\lambda_\mathsf{B}^-, \lambda_\mathsf{B}^+)$. Q.E.D.

Before proving Theorem 4, we first provide a constructive lemma and its proof.

LEMMA 5. *For $\lambda \in (\lambda_{min}, \lambda_{max})$, unique thresholds $b_\lambda^- \in (0,1)$ and $b_\lambda^+ \in (0,1)$ exist such that*

$$\lambda \mathbb{P}(\Theta_t = \mathsf{A} \,|\, S^H = +) + (1-\lambda)\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S^M = -, b_{t-1}) \geq r \;\Leftrightarrow b_{t-1} \leq b_\lambda^- \,, \tag{63}$$

$$\lambda \mathbb{P}(\Theta_t = \mathsf{A} \,|\, S^H = -) + (1-\lambda)\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S^M = +, b_{t-1}) \geq r \;\Leftrightarrow b_{t-1} \leq b_\lambda^+ \,. \tag{64}$$

Proof of Lemma 5. We obtain the thresholds by solving the following equations.

$$\lambda \mathbb{P}(\Theta_t = \mathsf{A} \,|\, S^\mathsf{H} = +) + (1-\lambda)\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S^\mathsf{M} = -, b_\lambda^-) = r \,, \tag{65}$$

$$\lambda \mathbb{P}(\Theta_t = \mathsf{A} \,|\, S^\mathsf{H} = -) + (1-\lambda)\mathbb{P}(\Theta_t = \mathsf{A} \,|\, S^\mathsf{M} = +, b_\lambda^+) = r \,. \tag{66}$$

The closed-form expressions are

$$b_\lambda^- = \frac{\varphi_\lambda^-(1-\alpha^\mathsf{W}) - \beta^\mathsf{W}}{\beta^\mathsf{B} - \beta^\mathsf{W} + \varphi_\lambda^-(\alpha^\mathsf{B} - \alpha^\mathsf{W})} \text{ and } b_\lambda^+ = \frac{\bar{\beta}^\mathsf{W} - \varphi_\lambda^+ \alpha^\mathsf{W}}{\varphi_\lambda^+(\alpha^\mathsf{B} - \alpha^\mathsf{W}) + \beta^\mathsf{B} - \beta^\mathsf{W}}$$

where

$$\varphi_\lambda^- = \frac{p}{\bar{p}}\left[\frac{(1-\lambda) - r + \lambda \mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^\mathsf{H} = +)}{r - \lambda \mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^\mathsf{H} = +)}\right] \text{ and } \varphi_\lambda^+ = \frac{p}{\bar{p}}\left[\frac{(1-\lambda) - r + \lambda \mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^\mathsf{H} = -)}{r - \lambda \mathbb{P}(\Theta_t = \mathsf{A} \,|\, S_t^\mathsf{H} = -)}\right].$$

Finally, we can similarly define $b_\lambda^\mathsf{H} = \min(b_\lambda^-, b_\lambda^+)$ and $b_\lambda^\mathsf{M} = \max(b_\lambda^-, b_\lambda^+)$ as in Corollary 1. Q.E.D.

Proof of Theorem 4. Note that Lemma 5 shows that there exist thresholds $b_\lambda^-, b_\lambda^+ \in (0,1)$ for $\lambda \in (\lambda_{min}, \lambda_{max})$ as in the case of Lemma 1. Thus, the exact same steps at which thresholds $b^-, b^+ \in (0,1)$ are replaced by $b_\lambda^-, b_\lambda^+ \in (0,1)$ in the proofs of Lemma 3, Theorems 2-3 will imply this result because the results in those theorems do not depend on the exact values of thresholds b^- and b^+, as long as they are interior to $(0,1)$. Q.E.D.

Proof of Theorem 5. Note that Lemma 1 also characterizes the DM's decision rule for the biased case. Nevertheless, the belief updating (6) is replaced with (7.2) with a bias term μ when the machine's prediction is not correct. Thus, we need to modify Lemma 3 slightly to incorporate this change. In particular, the updated log-likelihood ratio process $\tilde{L}_t$ is as follows.

Case 1: If $b^+ > b^-$, we have $\tilde{L}_t = \tilde{L}_{t-1} + \tilde{R}_t^{\mathsf{HM}}(\tilde{L}_{t-1})$ where

$$\tilde{R}_t^{\mathsf{HM}}(\tilde{L}_{t-1}) \triangleq \begin{cases} 1_{\{S_t^{\mathsf{M}}=+\}}\left[1_{\{\Theta_t=\mathsf{A}\}}\log\left(\frac{\alpha^{\mathsf{B}}}{\alpha^{\mathsf{W}}}\right) + 1_{\{\Theta_t=\mathsf{NA}\}}\mu\log\left(\frac{\bar{\beta}^{\mathsf{B}}}{\bar{\beta}^{\mathsf{W}}}\right)\right] & \text{if } \tilde{L}_{t-1} \geq L_h \\ 1_{\{S_t^{\mathsf{H}}=+,S_t^{\mathsf{M}}=+\}}\left[1_{\{\Theta_t=\mathsf{A}\}}\log\left(\frac{\alpha^{\mathsf{B}}}{\alpha^{\mathsf{W}}}\right) + 1_{\{\Theta_t=\mathsf{NA}\}}\mu\log\left(\frac{\bar{\beta}^{\mathsf{B}}}{\bar{\beta}^{\mathsf{W}}}\right)\right] & \text{if } L_h > \tilde{L}_{t-1} > L_\ell \\ 1_{\{S_t^{\mathsf{H}}=+\}}\left[\tilde{\upsilon}_1 + \tilde{\upsilon}_2\right] & \text{if } L_\ell \geq \tilde{L}_{t-1} \end{cases}$$

with $L_h \triangleq \log\left(\frac{b^+}{\bar{b}^+}\right)$, $L_\ell \triangleq \log\left(\frac{b^-}{\bar{b}^-}\right)$ and

$$\tilde{\upsilon}_1 \triangleq 1_{\{\Theta_t=\mathsf{A}\}}\left(1_{\{S_t^{\mathsf{M}}=-\}}\mu\log\left(\frac{\bar{\alpha}^{\mathsf{B}}}{\bar{\alpha}^{\mathsf{W}}}\right) + 1_{\{S_t^{\mathsf{M}}=+\}}\log\left(\frac{\alpha^{\mathsf{B}}}{\alpha^{\mathsf{W}}}\right)\right),$$

$$\tilde{\upsilon}_2 \triangleq 1_{\{\Theta_t=\mathsf{NA}\}}\left(1_{\{S_t^{\mathsf{M}}=-\}}\log\left(\frac{\beta^{\mathsf{B}}}{\beta^{\mathsf{W}}}\right) + 1_{\{S_t^{\mathsf{M}}=+\}}\mu\log\left(\frac{\bar{\beta}^{\mathsf{B}}}{\bar{\beta}^{\mathsf{W}}}\right)\right).$$

Case 2: If $b^+ \leq b^-$, we have $\tilde{L}_t = \tilde{L}_{t-1} + \tilde{R}_t^{\mathsf{HM}}(\tilde{L}_{t-1})$ where

$$\tilde{R}_t^{\mathsf{HM}}(\tilde{L}_{t-1}) \triangleq \begin{cases} 1_{\{S_t^{\mathsf{M}}=+\}}\left[1_{\{\Theta_t=\mathsf{A}\}}\log\left(\frac{\alpha^{\mathsf{B}}}{\alpha^{\mathsf{W}}}\right) + 1_{\{\Theta_t=\mathsf{NA}\}}\mu\log\left(\frac{\bar{\beta}^{\mathsf{B}}}{\bar{\beta}^{\mathsf{W}}}\right)\right] & \text{if } \tilde{L}_{t-1} > L_h \\ 1_{\{\Theta_t=\mathsf{A}\}}\tilde{\nu}_1 + 1_{\{\Theta_t=\mathsf{NA}\}}\tilde{\nu}_2 & \text{if } L_h \geq \tilde{L}_{t-1} \geq L_\ell \\ 1_{\{S_t^{\mathsf{H}}=+\}}\left[\tilde{\upsilon}_1 + \tilde{\upsilon}_2\right] & \text{if } L_\ell > \tilde{L}_{t-1} \end{cases} \tag{67}$$

with $L_h \triangleq \log\left(\frac{b^-}{\bar{b}^-}\right)$, $L_\ell \triangleq \log\left(\frac{b^+}{\bar{b}^+}\right)$ and

$$\tilde{\nu}_1 \triangleq 1_{\{S_t^{\mathsf{M}}=+\}}\log\left(\frac{\alpha^{\mathsf{B}}}{\alpha^{\mathsf{W}}}\right) + 1_{\{S_t^{\mathsf{M}}=-\}}1_{\{S_t^{\mathsf{H}}=+\}}\mu\log\left(\frac{\bar{\alpha}^{\mathsf{B}}}{\bar{\alpha}^{\mathsf{W}}}\right),$$

$$\tilde{\nu}_2 \triangleq 1_{\{S_t^{\mathsf{M}}=+\}}\mu\log\left(\frac{\bar{\beta}^{\mathsf{B}}}{\bar{\beta}^{\mathsf{W}}}\right) + 1_{\{S_t^{\mathsf{M}}=-\}}1_{\{S_t^{\mathsf{H}}=+\}}\log\left(\frac{\beta^{\mathsf{B}}}{\beta^{\mathsf{W}}}\right).$$

Note that the thresholds in the theorem are determined as the break-even points of the following equations.

$$p\alpha^{\mathsf{B}}\log\left(\frac{\alpha^{\mathsf{B}}}{\alpha^{\mathsf{W}}}\right) + \bar{p}\bar{\beta}^{\mathsf{B}}\mu^{\mathsf{B}}\log\left(\frac{\bar{\beta}^{\mathsf{B}}}{\bar{\beta}^{\mathsf{W}}}\right) = 0 \tag{68}$$

$$p\alpha^{\mathsf{W}}\log\left(\frac{\alpha^{\mathsf{B}}}{\alpha^{\mathsf{W}}}\right) + \bar{p}\bar{\beta}^{\mathsf{W}}\mu^{\mathsf{W}}\log\left(\frac{\bar{\beta}^{\mathsf{B}}}{\bar{\beta}^{\mathsf{W}}}\right) = 0 \tag{69}$$

$$p\alpha^{\mathsf{H}}\left(\bar{\alpha}^{\mathsf{B}}\mu^{\mathsf{H}}\log\left(\frac{\bar{\alpha}^{\mathsf{B}}}{\bar{\alpha}^{\mathsf{W}}}\right) + \alpha^{\mathsf{B}}\log\left(\frac{\alpha^{\mathsf{B}}}{\alpha^{\mathsf{W}}}\right)\right) + \bar{p}\bar{\beta}^{\mathsf{H}}\left(\beta^{\mathsf{B}}\log\left(\frac{\beta^{\mathsf{B}}}{\beta^{\mathsf{W}}}\right) + \bar{\beta}^{\mathsf{B}}\mu^{\mathsf{H}}\log\left(\frac{\bar{\beta}^{\mathsf{B}}}{\bar{\beta}^{\mathsf{W}}}\right)\right) = 0 \tag{70}$$

The left-hand sides of these equations correspond to the mean of $\tilde{L}_{t+1} - \tilde{L}_t$ when evaluated at μ in place of the thresholds at corresponding values of $\tilde{L}_t$. In the remainder of the proof, we explain the sign of the mean of $\tilde{L}_{t+1} - \tilde{L}_t$ given the machine's type and the value of μ. When the sign of this mean is known, the results follow from the proofs of previous theorems on which we elaborate.

- when $\Gamma = \mathsf{B}$,

  — $\mu \geq \mu^{\mathsf{B}}$ and $\mu > \mu^{\mathsf{H}}$ implies that the mean of $\tilde{L}_{t+1} - \tilde{L}_t$ for $L_\ell \geq \tilde{L}_t$ for Case 1 ($L_\ell > \tilde{L}_t$ for Case 2) is negative. Further, the mean of $\tilde{L}_{t+1} - \tilde{L}_t$ for $\tilde{L}_t \geq L_h$ for Case 1 ($\tilde{L}_t > L_h$ for Case 2) is nonpositive. Thus, the proof of the first part (for $p \leq p^{\mathsf{W}}$) of Theorem 3 applies to this parameter regime.

— $\mu^{\mathsf{B}} > \mu > \mu^{\mathsf{H}}$ implies that the mean of $\tilde{L}_{t+1} - \tilde{L}_t$ for $L_\ell \geq \tilde{L}_t$ for Case 1 ($L_\ell > \tilde{L}_t$ for Case 2) is positive. Further, the mean of $\tilde{L}_{t+1} - \tilde{L}_t$ for $\tilde{L}_t \geq L_h$ for Case 1 ($\tilde{L}_t > L_h$ for Case 2) is negative. Thus, the proof of the second part (for $p > p^{\mathsf{W}}$) of Theorem 3 applies to this parameter regime.

— $\mu^{\mathsf{H}} \geq \mu \geq \mu^{\mathsf{B}}$ implies that the mean of $\tilde{L}_{t+1} - \tilde{L}_t$ for $L_\ell \geq \tilde{L}_t$ for Case 1 ($L_\ell > \tilde{L}_t$ for Case 2) is nonnegative. Further, the mean of $\tilde{L}_{t+1} - \tilde{L}_t$ for $\tilde{L}_t \geq L_h$ for Case 1 ($\tilde{L}_t > L_h$ for Case 2) is nonpositive. Thus, the proof of the first part (for $p \leq p^{\mathsf{B}}$) of Theorem 2 applies to this parameter regime.

— if $\mu^{\mathsf{B}} > \mu$ and $\mu^{\mathsf{H}} \geq \mu$ implies that the mean of $\tilde{L}_{t+1} - \tilde{L}_t$ for $L_\ell \geq \tilde{L}_t$ for Case 1 ($L_\ell > \tilde{L}_t$ for Case 2) is nonnegative. Further, the mean of $\tilde{L}_{t+1} - \tilde{L}_t$ for $\tilde{L}_t \geq L_h$ for Case 1 ($\tilde{L}_t > L_h$ for Case 2) is positive. Thus, the proof of the second part (for $p > p^{\mathsf{B}}$) of Theorem 2 applies to this parameter regime.

- when $\Gamma = \mathsf{W}$,

— $\mu \geq \mu^{\mathsf{W}}$ implies that the mean of $\tilde{L}_{t+1} - \tilde{L}_t$ for $\tilde{L}_t \geq L_h$ for Case 1 ($\tilde{L}_t > L_h$ for Case 2) is nonpositive. Further, the mean of $\tilde{L}_{t+1} - \tilde{L}_t$ for $L_\ell \geq \tilde{L}_t$ for Case 1 ($L_\ell > \tilde{L}_t$ for Case 2) is negative for all $\mu \geq 1$. Thus, the proof of the first part (for $p \leq p^{\mathsf{W}}$) of Theorem 3 applies to this parameter regime.

— $\mu^{\mathsf{W}} > \mu$ implies that the mean of $\tilde{L}_{t+1} - \tilde{L}_t$ for $\tilde{L}_t \geq L_h$ for Case 1 ($\tilde{L}_t > L_h$ for Case 2) is positive. Further, the mean of $\tilde{L}_{t+1} - \tilde{L}_t$ for $L_\ell \geq \tilde{L}_t$ for Case 1 ($L_\ell > \tilde{L}_t$ for Case 2) is negative for all $\mu \geq 1$. Thus, the proof of the second part (for $p > p^{\mathsf{W}}$) of Theorem 3 applies to this parameter regime.

Finally, we show that $\mu^{\mathsf{H}} > 1$. Note that the following term is always positive when evaluated at $\mu^{\mathsf{H}} = 1$; see Footnote 6.

$$p\alpha^{\mathsf{H}}\left(\bar{\alpha}^{\mathsf{B}}\mu^{\mathsf{H}}\log\left(\frac{\bar{\alpha}^{\mathsf{B}}}{\bar{\alpha}^{\mathsf{W}}}\right) + \alpha^{\mathsf{B}}\log\left(\frac{\alpha^{\mathsf{B}}}{\alpha^{\mathsf{W}}}\right)\right) + \bar{p}\bar{\beta}^{\mathsf{H}}\left(\beta^{\mathsf{B}}\log\left(\frac{\beta^{\mathsf{B}}}{\beta^{\mathsf{W}}}\right) + \bar{\beta}^{\mathsf{B}}\mu^{\mathsf{H}}\log\left(\frac{\bar{\beta}^{\mathsf{B}}}{\bar{\beta}^{\mathsf{W}}}\right)\right)$$

To make this equal to 0, μ^{H} has to be larger than 1 because it is decreasing in μ^{H}. Hence, we conclude the proof. Q.E.D.

Recent ESMT Working Papers

	ESMT No.
Mismanaging diagnostic accuracy under congestion Mirko Kremer, Frankfurt School of Finance and Management Francis de Véricourt, ESMT Berlin	**22-01**
The economics of dependence: A theory of relativity Hans W. Friederiszick, ESMT Berlin and E.CA Economics Steffen Reinhold, E.CA Economics and ECARES –Université Libre de Bruxelles	**21-02**
Beyond retail stores: Managing product proliferation along the supply chain Işık Biçer, SchulichSchool of Business, York University Florian Lücker, Cass Business School, City, University of London Tamer Boyaci, ESMT Berlin	**19-02 (R3)**
Effectiveness and efficiency of state aid for new broadband networks: Evidence from OECD member states Wolfgang Briglauer, Vienna University of Economics and Business (WU) Michał Grajek, ESMT Berlin	**21-01**
Contracting, pricing, and data collection under the AI flywheel effect Huseyin Gurkan, ESMT Berlin Francis de Véricourt, ESMT Berlin	**20-01 (R3)**