CENTRE DE RECHERCHE
SUR LES STRATÉGIES ÉCONOMIQUES

Advisory algorithms, automation bias and liability rules

Yves Oytana, Marie Obidzinski

June 2025

Working paper No. 2025 – 08

CRESE
30, avenue de l'Observatoire
25009 Besançon
France
http://crese.univ-fcomte.fr/

UNIVERSITÉ MARIE & LOUIS PASTEUR

Advisory algorithms, automation bias and liability rules

Marie Obidzinski,* Yves Oytana†

June 6, 2025

Abstract

We study the design of optimal liability sharing rules when the use of an AI prediction by a human user may cause external damage. To do so, we set up a game-theoretic model in which an AI manufacturer chooses the level of accuracy with which an AI is developed (which increases the reliability of its prediction) and the price at which it is distributed. The user then decides whether to buy the AI. The AI's prediction gives a signal about the state of the world, while the user chooses her effort to discover the payoffs in each possible state of the world. The user may be susceptible to an automation bias that leads her to overestimate the algorithm's accuracy (overestimation bias). In the absence of an automation bias, we find that full user liability is optimal. However, when the user is prone to an overestimation bias, increasing the share of liability borne by the AI manufacturer can be beneficial for two reasons. First, it reduces the rent that the AI manufacturer can extract by exploiting the user's overestimation bias by underinvesting or overinvesting in the AI accuracy. Second, due to the nature of the interaction between algorithm accuracy and the user effort, the user may be incentivized to increase her (too low) judgment effort.

Keywords: liability sharing; advisory algorithm; automation bias; prediction; judgment effort

JEL classification: K13.

*Université Paris Panthéon Assas, CRED (EA 7321), 75005 Paris, France. E-mail: marie.obidzinski@assas-universite.fr

†Université Marie et Louis Pasteur, CRESE (UR 3190), F-25000 Besançon, France. E-mail: yves.oytana@univ-fcomte.fr

The authors thank Christian At, John Shahar Dillbary, Luigi Franzoni, Julien Jacob, Alain Marciano, Gerd Muehlheusser, Joshua C. Teitelbaum, and the participants at the 70th Congress of the French Economic Association annual conference, the 39th European Association in Law and Economics annual conference, the 19th German Law and Economics annual conference, the 7th French Law and Economics annual conference, and the seminar organized by the ZiF Research Group on Economic and Legal Challenges in the Advent of Smart Products at Bielefeld University.

1 Introduction

Motivation. As artificial intelligence (AI) gains momentum, algorithms are being extensively used in AI-assisted decision-making (Rastogi et al., 2020). *Advisory* algorithms provide decision support in a variety of situations, while *performative* algorithms are "able to accomplish independent actions by gathering information, decide and execute" (Jussupow et al., 2020).[1] These two types of algorithms (advisory versus performative) are distinguished by their degree of autonomy. For example, in aviation, pilots have been transformed from operators to supervisors as a result of increased reliance on performative algorithms. We can expect to see the same trend for drivers in automated cars in the coming years. Conversely, advisory algorithms leave the final decision to the user. The focus of this paper is on advisory algorithms, which are used in a variety of contexts. Physicians use them to better interpret x-rays pictures, to better predict the onset and/or evolution of a disease, and to better anticipate emerging infectious disease epidemics.[2] Advisory algorithms are used by judges to help them assess the risk of recidivism of criminal defendants (*e.g.*, the controversial COMPAS algorithm, for "Correctional Offender Management Profiling for Alternative Sanctions").[3] Banking institutions also use predictions of credit scoring models to reduce the risk of default or to prevent fraud.

Because AI algorithms often outperform humans in a wide range of applications, there are great benefits to be gained from their use. For example, AI algorithms outperform humans in the context of pretrial release decisions (Kleinberg et al., 2018) or in the context of medical imaging analyses such as computed tomography (Cheng et al., 2016).[4] By reducing the cost of providing

[1] This classification originates from Nissen (2001) and Nissen and Sengupta (2006).

[2] See, for example, Chopard and Musy (2022) who study the market for AI systems in health care, and Dai and Singh (2023) who focus on the decision of impurely altruistic physicians whether to use an AI and to follow its recommendation when that recommendation may affect the finding of negligence.

[3] COMPAS has been criticized on the grounds that the algorithm is allegedly biased against black defendants (see the propublica article by Larson, Mattu, Kirchner, and Angwin, 2016). However, we should keep in mind that these types of algorithms are generally less biased than human decision makers, which may be a motivation for their use (see, *e.g.*, Kleinberg et al., 2018).

[4] For example, focusing on bail decisions and comparing a machine learning algorithm to judges, Kleinberg et al. (2018) show that using the algorithm can lead to large welfare gains: "crime can be reduced by up to 24.8% with no change in jailing rates, or jail populations can be reduced by 42.0% with no increase in crime rates." Shulayeva et al. (2017) compares the ability to analyze existing case law by a human or an automated annotators, both trained by a professional annotator. The algorithm allows to reach a correct classification level of 85% *versus* 83.7% for the human annotator.

high accuracy predictions, AI enables human operators to "know more about their environment, including about future states of the world" (Agrawal et al., 2018) and thus to make better decisions.

However, even with the assistance of an advisory algorithm, human decisions are prone to error. This is especially true if the human operator does not properly consider the reliability of the algorithm's prediction. In fact, as is well known in the computer science literature, humans may be overly reliant on the predictions made (Zerilli et al., 2019; Springer et al., 2017). Different terminologies are used to characterize this issue, such as the "control problem" (Zerilli et al., 2019), or the "misuse" issue (Parasuraman and Riley, 1997).[5] According to Zerilli et al. (2019), the control problem refers to "the tendency of the human agent within a human-machine control loop to become complacent, over-reliant or unduly diffident when faced with the outputs of a reliable autonomous system." This tendency has been referred to as "automation bias", notably by Cummings (2017) and Mosier and Skitka (2018). In our paper, the control problem or "misuse" (hereafter referred to as *overestimation bias*) is characterized by an overestimation of the probability that the algorithm's prediction is correct.[6] The bias can lead the human operator to use the algorithm inappropriately by reducing her own effort and the AI manufacturer to underinvest or overinvest in the accuracy of the AI.

One way to reduce the risk of external damage is to implement accountability mechanisms to share the burden of poor outcomes from collaborative human-AI decisions. Well-designed legislation can incentivize the production of high-accuracy algorithms, while mitigating the effects of behavioral biases and ensuring the appropriate use of predictions by users. In fact, AI legal frameworks are currently under discussion in many countries.[7]

[5] Misuse is specifically defined by Parasuraman and Riley (1997) as overreliance on automation.

[6] The explainable AI approach tends to address the problem of overestimation bias by providing insight into how the algorithm makes its prediction. However, this approach has not been very successful in achieving this goal (Buçinca et al., 2021).

[7] Legislation could take the form of regulations (*e.g.*, certifications), defective product liability, or specific AI tort laws. In the European Union, the Product Liability Directive (PLD) was recently revised in October 2024, notably to include other types of goods, including software. However, the AI Liability Directive (AILD) proposal was withdrawn in February 2025. The AILD was intended to complement the Artificial Intelligence Act (Regulation 2024/1689). Furthermore, specific sector regulations may exist, such as those in healthcare (Chopard and Musy, 2024). Thus, depending on the context, different liability rules may apply.

Research questions. Our paper addresses the issue of the optimal liability rule when an AI manufacturer develops an algorithm to be used by a human operator. The accuracy of the prediction is chosen by the AI manufacturer during the algorithm development phase. Then, a human operator chooses whether to use an algorithm (*i.e.* to pay for a prediction), and her level of unobservable cognitive effort (which we call *judgment effort*) to learn about the *payoff*. Finally, a decision is made based on the available information about the state of the world (which may be imperfectly revealed by the use of the algorithm) and the associated payoffs (which may be observed by way of the judgment effort). In the absence of any cognitive bias, a strict liability of the human operator induces her to use the algorithm in an appropriate way and, moreover, incentivizes the AI manufacturer to make the socially optimal investment in the algorithm's accuracy (since he fully internalizes the expected liability cost through the price). We then ask the following question: Could the overestimation bias justify a sharing of liability between the user of the algorithm and the AI manufacturer?

Assumptions and main results. In our model, we assume that the user of an algorithm suffers from an overestimation bias, which leads to a misperception of the risk of a false prediction. More specifically, the user tends to overestimate the accuracy of the algorithm. As explained by Miceli and Segerson (2021) and following Zeiler (2019), this bias is a "psychological mistake"[8] that affects the actual decisions made by the user, with the consequence that these decisions do not reflect the true costs and benefits that they face. Thus, the decisions made will not necessarily be optimal for the user (who may regret them later) as well as for society. Miceli and Segerson (2021) suggest that this implies that "there is a potential role for legal rules to correct the distortions in decision-making that these biases can create." For this reason, we consider the legislative authority responsible for choosing the liability rule to be "paternalistic" rather than "populist" in the sense emphasized by Salanié and Treich (2009), meaning that this authority perceives welfare using the true probability of a false prediction, as opposed to the probability perceived by the user who suffers from an

[8] As these authors explain, another type of "bias" comes from non-standard preferences. Unlike a misperception bias, a non-standard preferences bias does not imply that the individual is making a mistake, and thus should not necessarily be corrected.

overestimation bias.

Following Agrawal et al. (2019b), the algorithm's prediction and the user's judgment effort cover two different dimensions. While the information provided by the prediction refers to the actual state of the world (*e.g.*, whether a patient has cancer or not), the judgment effort refers to the payoffs in each possible state of the world (*e.g.*, whether the patient will benefit from intensive and expensive treatment if he has cancer).[9]

Our main results are as follows. When the human operator does not suffer from an overestimation bias, strict user liability is optimal. Indeed, the expected social cost is fully internalized in the price of the algorithm, resulting in the socially optimal investment in the AI accuracy, and an optimal level of judgment effort.

In contrast, we show that strict user liability is not always optimal when the user suffers from an overestimation bias, mainly for two reasons. The first reason relates to the choice of AI accuracy by the manufacturer: the AI manufacturer exploits the user's misperception, which leads to a suboptimal accuracy level. Specifically, the accuracy level of the AI will be too high (low) if the user overestimates (underestimates) the marginal effect of a higher investment by the AI manufacturer. In general (although we show that this is not always true), increasing the AI manufacturer's share of liability will induce him to choose a accuracy level closer to the first-best. The second reason relates to the user's choice of judgment effort. In fact, the user's overreliance on the algorithm will lead her to choose too low a level of judgment effort. In certain conditions pertaining to the nature of the interaction between algorithm accuracy and user effort, extending the AI manufacturer's liability share may prompt the user to enhance her judgment effort.

The rest of the paper is organized as follows. Section 2 presents the related literature. Section 3 is the model setup. Section 4 gives the first-best optimum, and we solve the model and comment on

[9]The approach of AI we adopt is close to Agrawal et al. (2018), Agrawal et al. (2019a), and Agrawal et al. (2019b). In their setting, the human operator can make a judgment effort (which is a cognitive effort) to assess the payoff in each possible state of the world, while the AI, when used, provides her with a prediction about the actual state of the world. We borrow their modeling approach and add a potential external damage when a risky decision is made. Both the effort of judgment and the accuracy of the AI can affect the decision and thus the occurrence of damage.

the second-best liability sharing rule in Section 5. Finally, Section 6 concludes and discusses our results.

2 Related literature

Our paper lies at the intersection of two bodies of literature: one on human decision-making and predictive algorithms (subsection 2.1) and the other on the economic approach of product liability (subsection 2.2). To our knowledge, the theoretical law and economics literature does not address the specifics of human-algorithm interactions, such as automation bias, when analyzing the efficiency of different liability rules.[10]

2.1 Human decision making and predictive algorithms

Decision-making with predictive algorithms has been investigated theoretically in a series of papers by Agrawal et al. (2019a, 2018, 2019b). Specifically, the authors consider the complementarity between an algorithm's prediction and the judgment of a human decision-maker. Our approach is similar, except we assume that the human decision-maker may incorrectly estimate the algorithm's accuracy.

Several computer science and economics papers have identified the limits of using predictive algorithms empirically. This literature aims to determine whether AI improves decision-making and, if so, under what conditions (Alur et al., 2024). The effectiveness of human-AI interactions appears to depend on various factors, including algorithmic tuning and user experience (Inkpen et al., 2023). Studies have been conducted in various sectors, including social services and healthcare. In the social sector, Fogliato et al. (2022) examined the use of an algorithmic risk assessment tool by social workers in child protection. The authors found that although the social workers altered their decisions in the presence of the algorithm, they did not exhibit significant algorithmic aversion or automation bias. In contrast, in the healthcare sector, the results of Agarwal et al. (2023) suggest that radiologists underuse the information produced by predictive algorithms. Other articles sug-

[10] Apart from Obidzinski and Oytana (2024), where the human is assumed to have a limited attention.

gest that AI users may rely too heavily on predictions. For instance, Keding and Meissner (2021) show that managers tend to rely too heavily on predictions made by AI-based advisory systems because they "associate this advice with a higher level of process structure and perceive it as more trustworthy than human advice."

In our paper, we focus primarily on the issue of overreliance, while explaining in Section 6 how our model can address algorithmic aversion.

2.2 The economic approach of product liability

The law and economics literature has provided some insights into tort liability for products based on artificial intelligence. This literature generally assumes that technology and humans are substitute means of achieving a task rather than complementary ones.[11] There is specific law and economics literature on the optimal liability of smart products (and more specifically, self-driving cars) when accidents involving smart products may arise. More broadly, our paper is also related to the literature on product liability and consumer biases, as well as the literature on product liability and sequential care.

Autonomous vehicles (AVs) and liability rules. There is a growing body of literature on the liability rules that should apply to AVs (Shavell, 2020; Talley, 2019; De Chiara et al., 2021; Dawid and Muehlheusser, 2022; Guerra et al., 2022a,b). These papers differ with respect to (1) whether the probability of an accident can be affected by the vehicle's mileage, or the manufacturer's investment, (2) whether there is a mix of automated vehicles and human-driven vehicles, and (3) the type of liability rules envisioned.[12] Our framework differs in that we consider that the victim is a passive third party who has no control over the probability of the accident occurring. In addition, we do not specifically deal with AVs; rather, we focus more generally on advisory algorithms and

[11] With the exception of Chopard and Musy (2024).

[12] For example, Shavell (2020) considers a model in which all vehicles are autonomous and proposes the use of a new form of liability in which damages are paid to the state. Talley (2019) and De Chiara et al. (2021) develop a model in which only some (but not all) vehicles are autonomous. Dawid and Muehlheusser (2022), in the context of a dynamic model of product innovation calibrated to the U.S. auto market, study how liability rules may affect the emergence and the development of AV.

the distinction between human judgment and algorithmic prediction. Nevertheless, we share some important results with these papers. Like Shavell (2020) and Talley (2019), we find that a strict liability regime may be suboptimal. Moreover, similar to De Chiara et al. (2021) and Dawid and Muehlheusser (2022), we find that increasing the share of liability borne by the manufacturer may improve accuracy. However, to our knowledge, the emerging literature on AVs has not yet considered the fact that human users may be prone to cognitive biases when interacting with machines.

Product liability and consumer biases. Our paper is more broadly related to the product liability literature.[13] Hay and Spier (2005) show that it is optimal for the consumer, if fully solvent, to bear full liability for external damage. This result is consistent with our benchmark without user bias. When consumers are insolvent, a "residual-manufacturer liability", where the liability is shared between the manufacturer and the consumer, may be optimal. We also find that a sharing of liability between the manufacturer and the consumer (*i.e.*, the human operator in our context) may be optimal, though the reasons for this result differ.[14] In Hay and Spier (2005), the consumer *cannot* be strictly liable because he is insolvent, while in our model the human operator *should not* be strictly liable because then the AI manufacturer would benefit from the consumer's misperception, resulting in a suboptimal level of accuracy. A subset of the product liability literature has considered biased consumers.[15] The closest to our paper are Friehe et al. (2020) and Obidzinski and Oytana (2024). Friehe et al. (2020) compare liability rules when consumers are present-biased, while Obidzinski and Oytana (2024) consider the case where users exhibit behavioral inattention. We share their results that consumer bias may provide a rationale for sharing liability between the consumer and the (monopolistic) manufacturer. However, we differ in the specific bias we consider and in the nature of the decisions made by users.[16] In a broader account of the role of bias in economic models of law, Miceli and Segerson (2021) recently consider the case where

[13] Daughety and Reinganum (2013) and Geistfeld (2009) provide surveys of the product liability literature. See also the seminal paper by Landes and Posner (1985).

[14] In our framework, we assume that both the manufacturer and the user are fully solvent.

[15] On consumer misperception, see the seminal papers by Spence (1977) and Polinsky and Rogerson (1983).

[16] See also Baniak and Grajzl (2017), who consider the consequences of possible customer misperceptions about future usage, referred to as *projection bias*.

consumers misperceive their risk of damage in both a competitive and in a monopolistic setting. In the perfectly competitive setting, misperception implies that strict producer liability is optimal. In the monopolistic setting, strict producer liability is optimal when consumers overestimate the risk, while liability sharing may be optimal when consumers underestimate the risk, because it (partially) offsets the monopoly distortion on quantities. In contrast, we do not consider how liability rules may lead to underproduction or overproduction, since in our model a representative user buys at most one prediction. Thus, there is no heterogeneity in the user's willingness to pay for the prediction of the AI.

Sequential care. In our paper, the risk of damage can be mitigated by both the accuracy of the algorithm chosen by the manufacturer and by the effort of the human operator. This framework has similarities to sequential care models in the economic approach to tort law (Wittman, 1981; Shavell, 1983). However, the insights provided by these models cannot be directly applied to the context under study. This is due to the fact that, even if the user's effort decision intervenes after the accuracy of the algorithm is set, the judgment effort is unobservable and, as a consequence, cannot be subject to a negligence rule.[17]

3 The Model

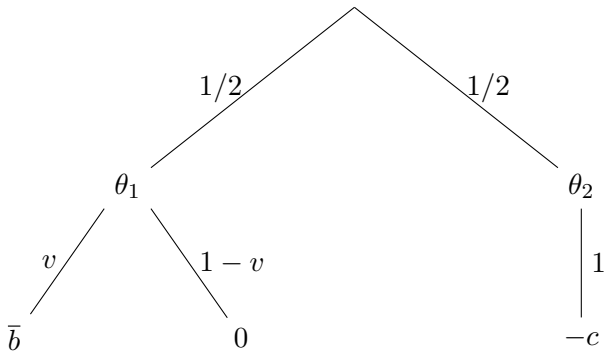
We build a model based on Agrawal et al. (2019b) (inspired by Bolton and Faure-Grimaud, 2009) to derive the optimal liability regime chosen by a policymaker when the use of an algorithm may cause external damage. The purpose of the liability regime is to apportion the damage between a representative user of the algorithm (the human operator, H, she) and the AI manufacturer (M, he), when the human operator may suffer from an overestimation bias. This bias leads the user of the algorithm to be overly confident in the prediction made by the AI. In this section, we introduce the general setup of our model by describing the players' objectives, their decision variables and their payoff. We also describe our behavioral assumption regarding automation bias. Finally, we

[17] Our contribution is also related to the literature on multiple injurers and the design of apportionment rules (Landes and Posner, 1980; Landes et al., 1987; Guttel et al., 2021; Ferey and Dehez, 2016; Kornhauser and Revesz, 1989). However, these papers are primarily concerned with dilution of liability and the risk of suboptimal care.

describe the sequence of events.

The state of the world is $\theta \in \{\theta_1, \theta_2\}$, with $\Pr(\theta = \theta_1) = \Pr(\theta = \theta_2) = 1/2$. In each of the states of the world, a "safe" or a "risky" decision can be made. A safe decision always yields a payoff of 0, regardless of the state of the world, while the payoff of a risky decision depends on the realization of the state of the world. In particular, a state $\theta = \theta_1$ captures a "good" state of the world in that a risky decision yields a possibly positive payoff $b \in \{\overline{b}, 0\}$, where b is a random variable with $\Pr(b = \overline{b}) = v$, $\Pr(b = 0) = 1 - v$, and $\overline{b} > 0$. In contrast, in the "bad" state of the world $\theta = \theta_2$, the payoff is always $-c < 0$ if a risky decision is made.

Figure 1: Social payoff of a risky decision



The AI manufacturer. The state of the world θ is unobservable. However, the AI manufacturer (M) can sell an AI that makes an (imperfect) prediction about θ, in the form of a signal $s \in \{\theta_1, \theta_2\}$. It is assumed that M is a monopolist whose objective is to maximize his expected profit by choosing the level of AI accuracy $q \in [1/2, 1]$, and the price at which the AI is sold to the user. The level of AI accuracy is the probability that the prediction is correct: $\Pr(s = \theta) = q$. The cost to the manufacturer of achieving an accuracy level q for his algorithm is denoted $c_M(q)$, with $c'_M(q) \geq 0$, $c''_M(q) > 0$, $c'_M(1/2) = 0$ and $\lim_{q \to 1} c'_M(q) = +\infty$.

The user. The human user (H) first decides whether or not to buy the AI (or, equivalently in our model, its prediction). H may be prone to an overestimation bias. In this case, the probability that the algorithm reveals the correct state of the world as perceived by the human operator is

$\tilde{q}(q) > q$. Increasing the accuracy level of the algorithm (q) has a positive effect on the accuracy as perceived by H ($\tilde{q}'(q) > 0$), although the magnitude of this increase may be underestimated ($\tilde{q}'(q) < 1$) or overestimated ($\tilde{q}'(q) > 1$).

After observing s, H exerts a judgment effort $e \in [0, 1]$. The judgment effort allows her to learn the payoff of a risky decision.[18] Specifically, with probability e, H observes the realization of b (*i.e.*, the payoff obtained from a risky decision if $\theta = \theta_1$). The judgment effort costs her $c_H(e)$, with $c'_H(e) \geq 0$, $c''_H(e) > 0$, $c'_H(0) = 0$ and $\lim_{e \to 1} c'_H(e) = +\infty$.

The user's objective is to minimize the sum of her liability cost, plus the cost of her effort and the price of the prediction.

The decision rule. We assume the following decision rule: *a favorable prediction $s = \theta_1$ is a necessary condition for making a risky decision. If $s = \theta_1$, then a risky decision is made, unless H's judgment effort reveals $b = 0$, in which case the safe decision is made.* In other words, a risky decision is made if and only if (i) the prediction is favorable and (ii) the operator does not learn bad news about the payoff in the "good" state of the world.

Note that we assume that the decision rule is fixed and cannot be changed by the user, even though this rule is not necessarily optimal for all parameters values. Nevertheless, we focus on this rule because it is relevant in the context of human-machine interaction, and because it allows for a much more tractable analysis compared to the case where H strategically chooses between a safe and a risky decision.

The policy maker. The policy maker chooses a liability rule to minimize the expected social cost, which is defined as the expected loss due to errors (making a safe decision when a risky one would have yielded a higher social benefit, and vice versa), plus the costs of algorithm accuracy $c_M(q)$ and judgment effort $c_H(e)$. We restrict our attention to a liability rule such that damages are

[18] This information is unavailable to the AI. As Alur et al. (2024) emphasise, "humans often have access to information which is not encoded in the data available to predictive algorithms."

shared between the AI manufacturer and the human operator. The liability sharing rule allocates a portion $\alpha \in [0,1]$ of the liability to the user of the algorithm. The AI manufacturer is liable for the remaining part $1-\alpha$. Thus, if $\alpha = 0$, the liability lies entirely with M (strict manufacturer liability), while if $\alpha = 1$, it lies entirely with H (strict user liability).

The timing of the game.

t=0. Nature independently chooses (i) the state of the world $\theta \in \{\theta_1, \theta_2\}$ and (ii) the payoff $b \in \{\bar{b}, 0\}$ of a risky decision in state $\theta = \theta_1$.

t=1. The policy maker chooses the liability sharing rule $\alpha \in [0,1]$.

t=2. The AI manufacturer chooses the AI accuracy $q \in [1/2, 1]$ and the price at which the prediction is sold.

t=3. The user decides whether to buy the prediction and observes a signal $s \in \{\theta_1, \theta_2\}$ about the state of the world.

t=4. The user chooses her level of judgment effort $e \in [0,1]$.

t=5. A decision (safe or risky) is made according to the decision rule given above, and payoffs are realized.

4 The first-best optimum

In this section, we derive the first-best optimum. In other words, we determine the levels of AI accuracy and judgment effort that minimize the expected social cost. This cost is defined as the sum of the total expected loss from errors (assuming no overestimation bias), the cost of accuracy, and the expected cost of judgment effort.

Suppose the AI prediction is $s = \theta_1$. This prediction is wrong with probability $(1-q)$. In this case, with probability $v + (1-v)(1-e)$, $b = 0$ is *not* observed and a risky decision is made (recall that the decision rule is such that a risky decision is made unless $b = 0$ is observed), causing a social

cost c. Conditional on a prediction $s = \theta_1$, the expected loss is:

$$L_{s=\theta_1}(q, e) = (1 - q)\Big(v + (1 - v)(1 - e)\Big)c \tag{1}$$

Now suppose the AI prediction is $s = \theta_2$. In this case, the decision rule is such that the safe decision is always made. However, with probability $(1 - q)$ this prediction does not match the state of the world: a risky decision would have yielded an expected payoff of $v\bar{b}$. Conditional on a prediction $s = \theta_2$, the expected loss is:

$$L_{s=\theta_2}(q) = (1 - q)v\bar{b} \tag{2}$$

We have $Pr(s = \theta_1) = Pr(s = \theta_2) = 1/2$. Thus, the total expected loss due to errors is:

$$L(q, e) = \frac{1}{2}L_{s=\theta_1}(q, e) + \frac{1}{2}L_{s=\theta_2}(q) \tag{3}$$

$L(q, e)$ is decreasing with the judgment effort (e):

$$\frac{\partial L}{\partial e}(q, e) = -\frac{1}{2}(1 - q)(1 - v)c < 0 \tag{4}$$

The probability of making a risky decision incorrectly decreases with the judgment effort. This is because, with probability e, the payoff b is observed by the user and, if $b = 0$, the "safe" decision is made, preventing any possible social cost c.

$L(q, e)$ is decreasing with the accuracy of the algorithm (q):

$$\frac{\partial L}{\partial q}(q, e) = -\frac{1}{2}\Big(v\bar{b} + \big((v + (1 - v)(1 - e)\big)c\Big) < 0 \tag{5}$$

There are two effects of increasing the accuracy of the AI (q) on the total expected loss. First, an increase in q improves the probability that a prediction $s = \theta_1$ will be correct, and thus the probability that a risky decision will be made when it actually brings a social benefit $\bar{b}$. Second, increasing q reduces the probability that the prediction $s = \theta_1$ will be incorrect, and thus the

probability that an incorrect risky decision will be made.

Note that a strictly positive judgment effort is only made if the prediction is $s = \theta_1$. Indeed, if $s = \theta_2$, the optimal level of judgment effort is 0, because increasing the judgment effort is costly while it cannot change the decision (which is always a safe decision). Thus, with a slight abuse of notation, we will refer to e as the level of judgment effort conditional on the prediction $s = \theta_1$ being observed.

The expected social cost is the sum of the total expected loss $L(q, e)$, plus the costs of algorithm accuracy ($c_M(q)$) and judgment effort ($c_H(e)$):

$$SC(q, e) = L(q, e) + c_M(q) + \frac{1}{2}c_H(e) \tag{6}$$

The first best level of AI accuracy (q), for a given e, minimizes the expected social cost and is characterized by the following first-order condition (FOC):

$$-\frac{\partial L}{\partial q}(q, e) = c'_M(q) \tag{7}$$

The socially optimal level of accuracy is reached when the marginal benefit of reducing the expected error loss (the left-hand side of (7)) is equal to the marginal cost of accuracy (the right-hand side of (7)).

Similarly, the first-best level of the judgment effort (e), for a given q, is characterized by the FOC:

$$-\frac{\partial L_{s=\theta_1}}{\partial e}(q, e) = c'_H(e) \tag{8}$$

The socially optimal judgment effort is achieved when the marginal benefit of reducing the expected error loss (the left-hand side of (8)) is equal to the marginal cost of the judgment effort (the right-hand side of (8)).

Now that we have determined the accuracy of the algorithm and the judgment at the first-best optimum, we consider the second-best in the next section.

5 Equilibrium judgment effort, accuracy and liability sharing rule

Recall that when the user suffers from an overestimation bias, the probability that the algorithm's prediction about the state of the world is correct, *as perceived by the human operator*, is $\tilde{q}(q) > q$ (and $\tilde{q}(q) = q$ when there is no bias). In this section, we solve the model, starting with the user's choice of her judgment effort both when she uses the AI and when she does not. We then derive her willingness to pay for the algorithm, the manufacturer's choice of algorithm accuracy, and finally the second-best liability sharing rule. At each step, we emphasize the effect of the user's overestimation bias.

5.1 The user's choice of her judgment effort

Case 1: with an AI. As emphasized in the previous section, when H uses an AI to obtain a prediction, an error may occur, generating a cost for which the user is liable for a part α (the manufacturer is liable for the remaining part $1-\alpha$). The user seeks to minimize her expected liability costs (keeping in mind that her perception of these costs will be inaccurate if she is prone to an overestimation bias), plus the cost of her judgment effort. Thus, her expected utility is given by:

$$\alpha L_{s=\theta_1}\big(\tilde{q}(q), e\big) + c_H(e) \tag{9}$$

Her choice of judgment effort is characterized by the FOC:

$$-\alpha \frac{\partial L_{s=\theta_1}}{\partial e}\big(\tilde{q}(q), e\big) = c'_H(e) \tag{10}$$

We denote the implicit solution of this FOC by $e^*_{AI}(\alpha, q)$. The interpretation of (10) is equivalent to (8) (*i.e.*, the user's judgment effort at the first-best optimum), with two differences: (i) the marginal social benefit (left-hand side of (10)) is weighted by $\alpha \leq 1$ (H internalizes only a fraction of the social benefit of her judgment effort if she is not fully liable), and (ii) if H overestimates

the accuracy of the AI, she underestimates the probability that the algorithm's prediction is wrong and thus her liability costs. These two differences negatively affect the user's judgment effort at equilibrium compared to the first-best. Thus, H's judgment effort is insufficient unless these two conditions are met: (i) H is not subject to an automation bias ($\tilde{q}(q) = q$) and (ii) she is fully liable for the social cost c in case of an erroneous risky decision ($\alpha = 1$).

Proposition 1. *For a given AI accuracy level q and if H is not subject to an overestimation bias ($\tilde{q}(q) = q$), H's judgment effort is socially optimal under strict user liability ($\alpha = 1$). If H has an overestimation bias ($\tilde{q}(q) > q$), her judgment effort is too low.*

Proof. If $\alpha = 1$ and $\tilde{q}(q) = q$, then (10) is equivalent to (8): for a given accuracy level q, the judgment effort chosen by the user is that which minimizes $SC(q, e)$. If $\tilde{q}(q) > q$, then for any $\alpha \in [0, 1]$, the left-hand term in (10) is less than the left-hand term in (8): the user's judgment effort is less than that which minimizes $SC(q, e)$. □

From the implicit function theorem, we can show that:

$$\frac{\partial e^*_{AI}}{\partial \alpha}(\alpha, q) = -\frac{\frac{\partial L_{s=\theta_1}}{\partial e}\big(\tilde{q}(q), e^*_{AI}(\alpha, q)\big)}{c''_H\big(e^*_{AI}(\alpha, q)\big)} > 0 \tag{11}$$

And:

$$\frac{\partial e^*_{AI}}{\partial q}(\alpha, q) = -\frac{\alpha\tilde{q}(q)\frac{\partial^2 L_{s=\theta_1}}{\partial e \partial q}\big(\tilde{q}(q), e^*_{AI}(\alpha, q)\big)}{c''_H\big(e^*_{AI}(\alpha, q)\big)} < (=)0 \text{ if } \alpha > (=)0 \tag{12}$$

From (11), increasing H's liability share (α) has a positive effect on her judgment effort. From (12), increasing the accuracy level of the AI reduces H's judgment effort if she bears at least part of the social cost of an erroneous risky decision ($\alpha > 0$). Otherwise (when $\alpha = 0$), H's judgment effort is zero, regardless of the accuracy of the AI.

Case 2: without AI. If H does *not* use the AI, we assume that she receives a signal about the state of the world that is correct ($s = \theta$) with probability 1/2. Since no AI is used, she bears full

liability in case of an error. According to the assumed decision rule, her expected perceived cost is:

$$L_{s=\theta_1}(1/2, e) + c_H(e) \tag{13}$$

Her choice of judgment effort is characterized by the FOC:

$$\frac{1}{2}(1-v)c = c'_H(e) \tag{14}$$

We denote the implicit solution of this FOC by $e^*_\emptyset$. A comparison of (10) and (14) shows that this judgment effort is higher than when using the AI. This is because the probability of error about the state of the world is higher, and H fully internalizes the social cost c, which incentivizes her to make a higher effort. Note that the liability sharing rule (α) and the accuracy of the AI (q) have no effect on the effort $e^*_\emptyset$, since the AI is not used.

5.2 The user's willingness to pay

When using the AI, the user's expected *perceived* cost of choosing a judgment effort $e^*_{AI}(\alpha, q)$ is:

$$C_{H,AI}(\alpha, q) = \alpha L(\tilde{q}(q), e^*_{AI}(\alpha, q)) + \frac{1}{2}c_H(e^*_{AI}(\alpha, q)) \tag{15}$$

Conversely, if she does not use the AI, her expected *perceived* cost of choosing a judgment effort $e^*_\emptyset$ is:

$$C_{H,\emptyset} = L(1/2, e^*_\emptyset) + \frac{1}{2}c_H(e^*_\emptyset) \tag{16}$$

The expected (perceived and actual) costs borne by H are higher when the algorithm is not used, for two reasons. First, without AI, the user's information about the state of the world is less reliable. Thus, the probability of error is higher than with the AI. Second, the manufacturer cannot be held liable for the cost of errors, because the AI is not used. Therefore, H bears the full expected social cost of errors. H's willingness to pay for the AI is:

$$P(\alpha, q) = C_{H,\emptyset} - C_{H,AI}(\alpha, q) > 0 \tag{17}$$

Using the envelope theorem, we find:

$$\frac{\partial P}{\partial \alpha}(\alpha, q) < 0 \tag{18}$$

And:

$$\frac{\partial P}{\partial q}(\alpha, q) > (=)0 \text{ if } \alpha > (=)0 \tag{19}$$

From (18), increasing the user's share of liability (or, equivalently, decreasing the manufacturer's share) reduces the user's willingness to pay. Indeed, one of the benefits of AI for the user is that it shifts some of the liability to the manufacturer. From (19), increasing the accuracy of the AI increases the user's willingness to pay because using the AI improves the reliability of the prediction and reduces the probability (and thus the expected cost) of errors, which the user internalizes (at least in part, if $\alpha > 0$) through liability.

5.3 The manufacturer's choice of AI accuracy

M's expected profit is equal to H's willingness to pay given by $P(\alpha, q)$,[19] minus his direct liability for losses due to errors $(1-\alpha)L(q, e^*_{AI}(\alpha, q))$ and the cost of investing in the accuracy of the AI $(c_M(q))$:

$$\Pi(\alpha, q) = P(\alpha, q) - (1-\alpha)L\big(q, e^*_{AI}(\alpha, q)\big) - c_M(q) \tag{20}$$

M's choice of AI accuracy is characterized by the FOC:

$$\frac{\partial P}{\partial q}(\alpha, q) - (1-\alpha)\frac{\partial L}{\partial q}(q, e^*_{AI}(\alpha, q)) - (1-\alpha)\frac{\partial L}{\partial e}(q, e^*_{AI}(\alpha, q))\frac{\partial e^*_{AI}}{\partial q}(\alpha, q) = c'_M(q) \tag{21}$$

The first (positive) term is the effect of AI accuracy on the user's willingness to pay. The second (positive) term is the direct effect of the AI accuracy (q) on the liability borne by M. The third (negative) term is the indirect effect of increasing the AI accuracy (q) on M's liability via its effect on H's judgment effort: as the AI accuracy increases, H's judgment effort decreases, which increases M's liability. To ensure an internal solution, we assume that the first two effects dominate the third.

[19] We assume that the manufacturer is a monopolist and is therefore able to extract all of H's surplus, since there is no heterogeneity in user preferences.

We denote the implicit solution of FOC (21) by $q^*(\alpha)$. Comparing (7) and (21), we find that the accuracy level chosen by M is socially optimal only if $\tilde{q}'(q) = 1$ (*e.g.*, if the user has no automation bias, with $\tilde{q}(q) = q$) and $\alpha = 1$ (strict user liability). First, the accuracy of the AI chosen by M is negatively affected by the fact that H's judgment effort decreases with q (recall that $\frac{\partial e^*_{AI}}{\partial q}(\alpha, q) \leq 0$). This reduction in judgment effort is costly to M because it increases the probability that an incorrect prediction $s = \theta_1$ will lead to a risky decision, generating an error whose loss is (at least partially) internalized by M if $\alpha < 1$. As a result, M will have an incentive to reduce the accuracy of the AI below the socially optimal level. Second, even if $\alpha = 1$, the accuracy level of the AI is too high if $\tilde{q}'(q^*(1)) > 1$ and too low if $\tilde{q}'(q^*(1)) < 1$. The intuition is as follows. When $\tilde{q}'\big(q^*(1)\big) > 1$, the accuracy level perceived by the user increases faster than the actual accuracy level. As a result, H overestimates the effect of higher accuracy on her liability. Since the user's (perceived) liability is internalized by M via the price, M has an incentive to choose an accuracy level that is too high (exploiting the user's overestimation bias). If $\tilde{q}'\big(q^*(1)\big) < 1$, the reasoning is similar: the effect of higher accuracy on the expected cost of errors is underestimated by H, and as a result the AI accuracy level chosen by M is too low.

Proposition 2. *If H is not subject to an overestimation bias ($\tilde{q}(q) = q$), then a regime of strict user liability ($\alpha = 1$) achieves the first-best optimum. If H is subject to an overestimation bias ($\tilde{q}(q) > q$) and is fully liable ($\alpha = 1$), then the accuracy level chosen by M is socially excessive (too low) if $\tilde{q}'\big(q^*(1)\big) > 1$ ($\tilde{q}'\big(q^*(1)\big) < 1$).*

Proof. Suppose that $\alpha = 1$. Substituting (10) into (21) gives us:

$$\frac{\partial P}{\partial q}(1, q) = c'_M(q) \tag{22}$$

With:

$$\frac{\partial P}{\partial q}(1, q) = -\tilde{q}'(q)\frac{\partial L}{\partial q}\big(q, e^*_{AI}(1, q)\big) \tag{23}$$

If $\tilde{q}(q) = q$, (22) is equivalent to (7) for $e = e^*_{AI}(1, q)$. Since $e^*_{AI}(1, q)$ is the first-best level of judgment effort for a given q according to Proposition 1 in this case, M chooses the first-best level

of AI accuracy.

If $\tilde{q}(q) > q$, then, according to (23), the left-hand side of (22) (*i.e.*, the marginal benefit of M from increasing the level of AI accuracy) is strictly higher than the left-hand side of (7) for $e = e^*_{AI}(1,q)$ (*i.e.*, the marginal social benefit from increasing the level of AI accuracy) if $\tilde{q}'(q^*(1)) > 1$. Conversely, the level of AI accuracy chosen by M is strictly lower than the first-best level of AI accuracy if $\tilde{q}'\big(q^*(1)\big) < 1$. □

The result of Proposition 2 shows that if H is subject to an overestimation bias, it is impossible to achieve the first-best, regardless of the liability sharing rule (unless we are in the special case $\tilde{q}'(q) = 1$).

Lemma 1. *Suppose H is fully liable ($\alpha = 1$) and subject to an overestimation bias ($\tilde{q}(q) > q$). If $\tilde{q}'(q) > 1$, increasing M's share of liability has a negative effect on the accuracy level of the AI ($q^{*\prime}(1) > 0$). If $\tilde{q}'(q) < 1$, the effect is ambiguous.*

Proof. Applying the implicit function theorem to the FOC (21), we find that the sign of $q^{*\prime}(1)$ is the same as the sign of:

$$\big(1 - \tilde{q}'\big(q^*(1)\big)\big)\,\frac{\partial L}{\partial q}\Big(q^*(1), e^*_{AI}\big(1, q^*(1)\big)\Big) - \frac{1}{2}\left[\frac{\partial e^*_{AI}}{\partial \alpha}(1, q^*(1))\tilde{q}'\big(q^*(1)\big) + \frac{\partial e^*_{AI}}{\partial q}(1, q^*(1))\big(1 - q^*(1)\big)\right](1-v)c \quad (24)$$

Using (11) and (12), the term in square brackets can be rewritten:

$$-\frac{\tilde{q}'\big(q^*(1)\big)}{c''_h\Big(e^*_{AI}\big(1, q^*(1)\big)\Big)}\Big(\tilde{q}\big(q^*(1)\big) - q^*(1)\Big)(1-v)c < 0 \quad (25)$$

Thus, if $\tilde{q}'\big(q^*(1)\big) > 1$, the first and second terms in (24) are both positive: increasing M's share of liability (decreasing α) decreases the AI accuracy chosen by M (*i.e.*, $q^{*\prime}(1) > 0$). Conversely, if $\tilde{q}'\big(q^*(1)\big) < 1$, the first term is now negative, while the second is positive: increasing M's share of liability (decreasing α) has an ambiguous effect on the AI accuracy chosen by M: the sign of $q^{*\prime}(1)$

is ambiguous. □

This lemma shows that increasing the liability share of M does not necessarily bring the manufacturer's choice of AI accuracy closer to the socially optimal accuracy. In particular, when $\tilde{q}'\big(q^*(1)\big) < 1$, there are two opposing effects. On the one hand, increasing the manufacturer's liability incentivizes the manufacturer to increase the AI accuracy, which is too low (see Proposition 2). On the other hand, increasing the AI accuracy will induce the user to reduce her judgment effort, thereby increasing the expected liability cost of M.

5.4 The second-best liability sharing rule

There exists a liability sharing rule such that M bears a part of the liability ($\alpha < 1$) which is better than full user liability ($\alpha = 1$) if:

$$\left.\frac{\mathrm{d}SC\Big(q^*(\alpha), e^*_{AI}\big(\alpha, q^*(\alpha)\big)\Big)}{\mathrm{d}\alpha}\right|_{\alpha=1} > 0 \tag{26}$$

Using the envelope theorem and rearranging, (26) can be rewritten as:

$$\begin{aligned}&\frac{\partial L}{\partial q}\Big(q^*(1), e^*_{AI}(1, q^*(1))\Big)q^{*\prime}(1)\Big(1 - \tilde{q}'(q^*(1))\Big)\\ &\quad + \frac{\partial L}{\partial e}\Big(q^*(1), e^*_{AI}(1, q^*(1))\Big)\frac{\tilde{q}\big(q^*(1)\big) - q^*(1)}{1 - q^*(1)}\left[\frac{\partial e^*_{AI}}{\partial \alpha}\big(1, q^*(1)\big) + q^{*\prime}(1)\frac{\partial e^*_{AI}}{\partial q}\big(1, q^*(1)\big)\right] < 0\end{aligned} \tag{27}$$

Proposition 3. *(i) If H is not subject to an overestimation bias ($\tilde{q}(q) = q$), then full user liability ($\alpha = 1$) achieves the first-best in terms of AI accuracy and judgment effort. (ii) If H is subject to an overestimation bias ($\tilde{q}(q) > q$), then there exists a liability sharing rule $\alpha < 1$ such that the second-best is achieved if (27) holds.*

Proof. Part (i) is a direct consequence of Proposition 1 and Proposition 2. The proof of part (ii) is in the text. □

Part (i) of Proposition 3 follows directly from Propositions 1 and 2. However, if H is subject to an

overestimation bias (part (ii) of Proposition 3), then understanding the effects at play requires an interpretation of (27).

The first term in (27) captures the effect of a change in the liability sharing rule on the expected social cost via M's choice of AI accuracy. As the liability share of M increases, this effect helps to reduce the expected social cost if:

$$q^{*\prime}(1)\Big(1-\tilde{q}'(q^*(1))\Big)<0 \tag{28}$$

The signs of $q^{*\prime}(1)$ and $1-\tilde{q}'\big(q^*(1)\big)$ are ambiguous. To interpret (28), let us first assume that $\tilde{q}'\big(q^*(1)\big)>1$. In this case, Proposition 2 and Lemma 1 tell us that, starting from a strict user liability rule ($\alpha=1$), the accuracy level of the AI chosen by M is excessive, but can be reduced by assigning him a positive share of the liability ($q^{*\prime}(1)>0$). In other words, condition (28) is satisfied: increasing M's liability share brings the accuracy level of the AI chosen by M closer to the socially optimal accuracy level of the AI.

If $\tilde{q}'\big(q^*(1)\big)<1$, then condition (28) may not hold. In fact, according to Proposition 1, the accuracy level chosen by M is too low, while the effect of increasing M's share of liability on his choice of AI accuracy is ambiguous. Under these conditions, it is possible that increasing M's share of liability will further decrease the AI accuracy, thereby increasing the expected social cost (the effort chosen by M is further away from the socially optimal effort).

The second term in (28) captures the effect of a change in the liability sharing rule on the expected social cost via H's judgment effort. As the liability share of M increases, this effect helps to reduce the expected social cost if:

$$\frac{\partial e^*_{AI}}{\partial\alpha}\big(1,q^*(1)\big)+q^{*\prime}(1)\frac{\partial e^*_{AI}}{\partial q}\big(1,q^*(1)\big)<0 \tag{29}$$

Suppose that $\tilde{q}'\big(q^*(1)\big)>1$. Again, from Proposition 2 and Lemma 1, starting from a strict user

liability rule ($\alpha = 1$), the accuracy level of the AI chosen by M is too high, but it can be reduced by giving him a positive share of the liability ($q^{*\prime}(1) > 0$). From Proposition 1, the judgment effort is too low and from (11) and (12) we know that $\frac{\partial e^*_{AI}}{\partial \alpha}\big(1, q^*(1)\big) > 0$ and $\frac{\partial e^*_{AI}}{\partial q}\big(1, q^*(1)\big) < 0$. There are two countervailing effects. First, a higher liability for M implies a lower liability for H, which reduces her already too low judgment effort (first term of (29)). Second, a larger share of liability attributed to M will lead to a reduction in the accuracy level of the AI chosen by M, causing H to increase her judgment effort (second term of (29)). If the second effect dominates the first, then condition (29) is satisfied: increasing M's share of liability has an overall positive effect on H's judgment effort, reducing the expected social cost.

If $\tilde{q}'\big(q^*(1)\big) < 1$, then the sign of $q^{*\prime}(1)$ is ambiguous. A higher liability for M still has a negative effect by incentivizing H to choose a lower judgment effort (first term of (29)). However, the indirect effect (second term of (29)), via M's choice of AI accuracy, is ambiguous: H's judgment effort increases if an increase in M's liability share causes M to choose a lower AI accuracy (*i.e.*, if $q^{*\prime}(1) > 0$), and vice versa.

Note that the magnitude of the effect captured by the second term in (27) (the effect of α on the expected social cost via H's judgment effort) increases with $\frac{\tilde{q}\big(q^*(1)\big) - q^*(1)}{1 - q^*(1)}$. This fraction is the magnitude of the overestimation bias ($\tilde{q}\big(q^*(1)\big) - q^*(1)$) relative to the probability of an incorrect prediction ($1 - q^*(1)$). It suggests that the effect captured by the second term in (27) tends to be stronger when users suffer from a large overestimation bias and use a very reliable AI, making the *relative* overestimation bias (*i.e.*, the percentage increase in accuracy as perceived by the user compared to the true accuracy) very high.

6 Discussion and conclusion

In this paper, we examine a situation in which human decision-maker and an AI advisory algorithm provide complementary information, with an incorrect decision resulting in damage for a third party. The accuracy of the algorithm and the user's judgment effort both impact the probability of an

incorrect decision, wich results in loss (*e.g.*, a physician implements a treatment that is not suited to the patient's condition after observing computed tomography and making an effort to assess the extent to which the treatment would benefit the patient). Following Agrawal et al., 2019a, 2018, 2019b, we assumed that the AI prediction and the human user's judgment effort are complementary because they provide information on separate dimensions. Specifically, the AI predicition provides information about the state of the world (*e.g.*, whether the patient is sick), while the user's judgment provides information about the payoff of a risky decision (*e.g.*, the payoff of implementing treatment for an ill patient). Furthermore, we assumed that the human user may be susceptible to automation bias (*i.e.*, she may overestimate the algorithm's accuracy). The objective of society is to minimize expected social costs, which are the sum of expected error costs and the costs of algorithm accuracy and judgment effort. In this context, we examine the optimal sharing of liability between the AI manufacturer and the human.

We show that, in the absence of overestimation bias, full user liability is optimal. However, when users are prone to an overestimation bias, increasing the AI manufacturer's liability can be beneficial. This reduces the rent that the AI manufacturer can extract by exploiting the user's overestimation bias through underinvestment or overinvestment in AI accuracy. Furthermore, due to the nature of the interaction between algorithm accuracy and the user effort, the user may be incentivized to increase her (too low) judgment effort.

In this concluding section, we discuss three possible extensions of our setting. First, we discuss algorithm aversion. Second, we examine another type of liability rule, namely the negligence rule. Third, we consider the possibility of debiasing the consumer by either the manufacturer or a public authority.

6.1 Algorithm aversion

Although we have focused on automation bias, users may be prone to other biases when using an AI. One such bias is algorithm aversion. Jussupow et al. (2020) define algorithm aversion as a "biased assessment of an algorithm which manifests in negative behaviors and attitudes towards the

algorithm compared to a human agent."[20] Our theoretical model can be used to assess how liability rules can mitigate the effects of users' algorithm aversion. In our setting, algorithm aversion can manifest itself in two ways.

First, algorithm aversion may reduce the user's willingness to pay for the prediction (*e.g.*, the user suffers a fixed disutility when choosing to use an AI). If this effect is strong enough, the AI manufacturer may choose not to develop the AI at all, even though its development and adoption would reduce the expected social cost.

Second, the user may underestimate the accuracy of the AI prediction as a result of algorithm aversion. In this case, we can reinterpret the results of our model with the alternative assumption that when the user is algorithm averse, the perceived AI accuracy is *inferior* to the true AI accuracy level. We can then expect the following results. In the absence of algorithm version, strict user liability allows the first-best optimum to be achieved. Now assume a strict liability of the user. Re-interpreting Proposition 1, we find that the user effort is now too high. The result of Proposition 2 holds in the sense that if the user overestimates (underestimates) the effect of an accuracy increase, then the AI accuracy chosen by the manufacturer is socially excessive (too low). Finally, we can find a condition analogous to the one in Proposition 3, under which it is socially beneficial to assign a strictly positive share of liability to the AI manufacturer. If the user is averse to algorithms, there are still two main effects. The first (second) captures the effect of increasing the manufacturer's liability share on the expected social loss via his choice of AI accuracy (the user's choice of judgment effort). Overall, the signs of these two effects will remain ambiguous and will largely depend on whether the user overestimates or underestimates the effect of an accuracy increase, as in our analysis of the overestimation bias.

[20] Note, however, that the evidence for the existence of algorithm aversion is mixed, especially with respect to users of advisory algorithms (Jussupow et al., 2020). Interestingly, in the health care context, Longoni et al. (2019) show that a specific form of algorithm aversion, which they call "uniqueness neglect", is eliminated when the AI makes a prediction that "only supports, rather than replaces, a decision made by a human healthcare provider", which is by definition the case with advisory algorithms.

6.2 Negligence rule

In our analysis, we have focused exclusively on liability sharing rules. However, other liability rules may be relevant in the same setting. One possible alternative is to establish a negligence rule that specifies a minimum level of accuracy for the AI. The AI manufacturer would be liable if this standard is not met, while the user would be liable otherwise. Although a well-designed negligence rule can avoid some of the limitations of a liability sharing rule, it may be difficult to implement in practice, especially when applied to external damages resulting from an incorrect AI prediction.[21]

A first limitation that may prevent a negligence rule from being fully efficient is that the socially optimal level of the standard may be difficult to determine. In our setting, it would require knowing both the effect of a greater investment in accuracy on the reliability of the algorithm's prediction, and the costs and benefits of increasing that accuracy. However, as Hay and Spier (1997) explain, "manufacturers are likely to be better informed about the feasibility of product modifications than regulators".

A second limitation is that the AI manufacturer may misperceive the level of the standard that will be enforced by the courts, either because of the vagueness of the terms used to formulate the standard or because of uncertainty about how the court will interpret that formulation. This is especially true if the standard can only be stated in very broad terms, which is likely to be the case because of the practical impossibility of specifying a precise level of accuracy in a fast-moving and complex technological environment.

A third limitation is that it may be costly for the court to observe how much investment the AI manufacturer actually made in improving the algorithm, and thus to determine accurately whether the manufacturer was negligent or not.[22] This is particularly true when we consider, for example, the predictions made by deep learning algorithms. In fact, these algorithms are often considered as

[21] For a general discussion of liability rules and AI, see Buiten et al. (2021).

[22] As is well known from the law and economics literature (*e.g.* Shavell, 1987), errors in determining negligence will often lead to a level of precaution that is higher than the socially optimal level. In the context of our model, this means that the AI manufacturer will choose an excessively high level of accuracy.

"black boxes" because it can be very difficult, even for their creators, to identify the weight given to each feature (a measurable property or characteristic of a data set) and how these features relate to each other to shape the algorithm's prediction.

Interestingly, our model points to a fourth limitation, which is a consequence of the user's overestimation bias. According to our results, if the level of AI accuracy *as perceived by the user* increases faster than the true accuracy (*i.e.*, the user overestimates the increase in accuracy), then the level of accuracy chosen by the AI manufacturer is too high. Since a standard specifies a minimum level of accuracy, a negligence rule will fail to restore the proper incentives of the manufacturer with respect to the accuracy of the AI.

Although these limitations may make it difficult to efficiently implement a negligence rule in practice, the applicability of a liability sharing rule is not straightforward either. Designing a very specific allocation of liability (*e.g.*, the sharing that minimizes the expected social cost when strict user liability is not second-best) can be challenging. Indeed, the optimal sharing rule will generally depend on the particular context in which it is applied. Although there will always be some uncertainty about the exact liability sharing that should apply, our analysis shows that when choosing how to share liability, the policymaker should consider not only the existence of a possible user overestimation bias, but also, among other things, the relative magnitude of the overestimation bias, whether the "subjective" AI accuracy (the accuracy as perceived by the user) is increasing faster than the "objective" AI accuracy (the true AI accuracy), and so on. Unfortunately, this information is likely to vary depending on the specific algorithms used and the context in which they are used.

6.3 Debiasing the consumer

So far, we have omitted both the possibility that the public authority may try debiasing the human user (Jolls and Sunstein, 2006; Luppi and Parisi, 2016), and the possibility that the manufacturer may want to educate the user (Bienenstock, 2016). In the latter case, when applied to our context of decision-making with an advisory algorithm, the user tends to overestimate the accuracy of the

product. However, it is not rational for a monopoly to educate the consumer, but it might be the case in an oligopoly context: a manufacturer might be willing to correct the consumer's perception of the accuracy of its competitors' products. Such an extension would be relevant in our setting, where it would be worth investigating how liability rules might induce AI manufacturers to invest in consumer education. These extensions are left for future research.

References

Agarwal, N., Moehring, A., Rajpurkar, P., and Salz, T. (2023). Combining human expertise with artificial intelligence: Experimental evidence from radiology. Technical report, National Bureau of Economic Research.

Agrawal, A., Gans, J., and Goldfarb, A. (2019a). Prediction, judgment, and complexity: a theory of decision-making and artificial intelligence. In *The economics of artificial intelligence: An agenda*, pages 89–110. University of Chicago Press.

Agrawal, A., Gans, J. S., and Goldfarb, A. (2018). Human judgment and ai pricing. In *AEA Papers and Proceedings*, volume 108, pages 58–63.

Agrawal, A., Gans, J. S., and Goldfarb, A. (2019b). Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy*, 47:1–6.

Alur, R., Laine, L., Li, D. K., Shung, D., Raghavan, M., and Shah, D. (2024). Integrating expert judgment and algorithmic decision making: An indistinguishability framework. *arXiv preprint arXiv:2410.08783*.

Baniak, A. and Grajzl, P. (2017). Optimal liability when consumers mispredict product usage. *American law and economics review*, 19(1):202–243.

Bienenstock, S. (2016). Consumer education: why the market doesn't work. *European Journal of Law and Economics*, 42:237–262.

Bolton, P. and Faure-Grimaud, A. (2009). Thinking ahead: the decision problem. *The Review of Economic Studies*, 76(4):1205–1238.

Buçinca, Z., Malaya, M. B., and Gajos, K. Z. (2021). To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21.

Buiten, M., de Streel, A., and Peitz, M. (2021). An economic approach to regulating algorithms. Technical report, Centre on Regulation in Europe.

Cheng, J.-Z., Ni, D., Chou, Y.-H., Qin, J., Tiu, C.-M., Chang, Y.-C., Huang, C.-S., Shen, D., and Chen, C.-M. (2016). Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in ct scans. *Scientific reports*, 6(1):1–13.

Chopard, B. and Musy, O. (2022). Market for artificial intelligence in health care and compensation for medical errors. In *MPRA paper*.

Chopard, B. and Musy, O. (2024). Optimal liability rules for combined human-ai health care decisions. *Available at SSRN 4765482*.

Cummings, M. L. (2017). *Automation bias in intelligent time critical decision support systems*. Routledge.

Dai, T. and Singh, S. (2023). Artificial intelligence on call: The physician's decision of whether to use ai in clinical practice. Working paper.

Daughety, A. F. and Reinganum, J. F. (2013). Economic analysis of products liability: theory. In *Research handbook on the economics of torts*. Edward Elgar Publishing.

Dawid, H. and Muehlheusser, G. (2022). Smart products: Liability, investments in product safety, and the timing of market introduction. *Journal of Economic Dynamics and Control*, 134.

De Chiara, A., Elizalde, I., Manna, E., and Segura-Moreiras, A. (2021). Car accidents in the age of robots. *International Review of Law and Economics*, 68:106–022.

Ferey, S. and Dehez, P. (2016). Multiple causation, apportionment, and the shapley value. *The Journal of Legal Studies*, 45(1):143–171.

Fogliato, R., De-Arteaga, M., and Chouldechova, A. (2022). A case for humans-in-the-loop: Decisions in the presence of misestimated algorithmic scores. *Available at SSRN*.

Friehe, T., Rößler, C., and Dong, X. (2020). Liability for third-party harm when harm-inflicting consumers are present biased. *American Law and Economics Review*, 22(1):75–104.

Geistfeld, M. A. (2009). Products liability. In *Encyclopedia of Law and Economics*. Edward Elgar Publishing Limited.

Guerra, A., Parisi, F., and Pi, D. (2022a). Liability for robots i: legal challenges. *Journal of Institutional Economics*, 18(3):331–343.

Guerra, A., Parisi, F., and Pi, D. (2022b). Liability for robots ii: an economic analysis. *Journal of Institutional Economics*, 18(4):553–568.

Guttel, E., Procaccia, Y., and Winter, E. (2021). Shared liability and excessive care. *The Journal of Law, Economics, and Organization*.

Hay, B. and Spier, K. E. (1997). Burdens of proof in civil litigation: An economic perspective. *Journal of Legal Studies*, 26(26):413–431.

Hay, B. and Spier, K. E. (2005). Manufacturer liability for harms caused by consumers to others. *American Economic Review*, 95(5):1700–1711.

Inkpen, K., Chappidi, S., Mallari, K., Nushi, B., Ramesh, D., Michelucci, P., Mandava, V., Vepřek, L. H., and Quinn, G. (2023). Advancing human-ai complementarity: The impact of user expertise and algorithmic tuning on joint decision making. *ACM Transactions on Computer-Human Interaction*, 30(5):1–29.

Jolls, C. and Sunstein, C. R. (2006). Debiasing through law. *The Journal of Legal Studies*, 35(1):199–242.

Jussupow, E., Benbasat, I., and Heinzl, A. (2020). Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion. ECIS 2020 Research Papers.

Keding, C. and Meissner, P. (2021). Managerial overreliance on ai-augmented decision-making processes: How the use of ai-based advisory systems shapes choice behavior in r&d investment decisions. *Technological Forecasting and Social Change*, 171:120970.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293.

Kornhauser, L. A. and Revesz, R. L. (1989). Sharing damages among multiple tortfeasors. *The Yale Law Journal*, 98(5):831–884.

Landes, W. M. and Posner, R. A. (1980). Joint and multiple tortfeasors: An economic analysis. *The Journal of Legal Studies*, 9(3):517–555.

Landes, W. M. and Posner, R. A. (1985). A positive economic analysis of products liability. *The Journal of Legal Studies*, 14(3):535–567.

Landes, W. M., Posner, R. A., et al. (1987). *The Economic Structure of Tort Law*. Harvard University Press.

Longoni, C., Bonezzi, A., and Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46:629–650.

Luppi, B. and Parisi, F. (2016). Optimal liability for optimistic tortfeasors. *European journal of law and economics*, 41(3):559–574.

Miceli, T. J. and Segerson, K. (2021). The role of bias in economic models of law. *Review of Law & Economics*.

Mosier, K. L. and Skitka, L. J. (2018). Human decision makers and automated decision aids: Made for each other? In *Automation and human performance: Theory and applications*, pages 201–220. CRC Press.

Nissen, M. E. (2001). Agent-based supply chain integration. *Information Technology and Management*, 2:289–312.

Nissen, M. E. and Sengupta, K. (2006). Incorporating software agents into supply chains: Experimental investigation with a procurement task. *Mis Quarterly*, pages 145–166.

Obidzinski, M. and Oytana, Y. (2024). Artificial intelligence, inattention and liability rules. *International Review of Law and Economics*, 79:106211.

Parasuraman, R. and Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253.

Polinsky, A. M. and Rogerson, W. P. (1983). Products liability, consumer misperceptions, and market power. *The Bell Journal of Economics*, 14(2):581–589.

Rastogi, C., Zhang, Y., Wei, D., Varshney, K. R., Dhurandhar, A., and Tomsett, R. (2020). Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *arXiv preprint arXiv:2010.07938*.

Salanié, F. and Treich, N. (2009). Regulation in happyville. *The Economic Journal*, 119(537):665–679.

Shavell, S. (1983). Torts in which victim and injurer act sequentially. *The Journal of Law and Economics*, 26(3):589–612.

Shavell, S. (1987). *Economic Analysis of Accident Law*. Harvard University Press.

Shavell, S. (2020). On the redesign of accident liability for the world of autonomous vehicles. *The Journal of Legal Studies*, 49(2):243–285.

Shulayeva, O., Siddharthan, A., and Wyner, A. (2017). Recognizing cited facts and principles in legal judgements. *Artificial Intelligence and Law*, 25:107–126.

Spence, M. (1977). Consumer misperceptions, product failure and producer liability. *Review of Economic Studies*, 44:561–572.

Springer, A., Hollis, V., and Whittaker, S. (2017). Dice in the black box: User experiences with an inscrutable algorithm. In *2017 AAAI Spring Symposium Series*.

Talley, E. (2019). Automatorts: How should accident law adapt to autonomous vehicles? lessons from law and economics.

Wittman, D. (1981). Optimal pricing of sequential inputs: Last clear chance, mitigation of damages, and related doctrines in the law. *The Journal of Legal Studies*, 10(1):65–91.

Zeiler, K. (2019). Mistaken about mistakes. *European Journal of Law and Economics*, 48:9–27.

Zerilli, J., Knott, A., Maclaurin, J., and Gavaghan, C. (2019). Algorithmic decision-making and the control problem. *Minds and Machines*, 29(4):555–578.