Advisory algorithms and liability rules

Marie Obidzinski, Yves Oytana

CENTRE DE RECHERCHE
SUR LES STRATÉGIES ÉCONOMIQUES

Advisory algorithms and liability rules

Marie Obidzinski, Yves Oytana

September 2022

Working paper No. 2022–04

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

UFR SJEPG
Sciences juridiques économiques politiques et de gestion
UNIVERSITÉ DE FRANCHE-COMTÉ

Advisory algorithms and liability rules

Marie Obidzinski,* Yves Oytana†

September 13, 2022

Abstract

We study the design of optimal liability rules when the use of an advisory algorithm by a human operator (she) may generate an external harm. An artificial intelligence (AI) manufacturer (he) chooses the level of quality with which the algorithm is developed and the price at which it is distributed. The AI makes a prediction about the state of the world to the human operator who buys it, who can then decide to exert a judgment effort to learn the payoffs in each possible state of the world. We show that when the human operator overestimates the algorithm's accuracy (overestimation bias), imposing a strict liability rule on her is not optimal, because the AI manufacturer will exploit the bias by under-investing in the quality of the algorithm. Conversely, imposing a strict liability rule on the AI manufacturer may not be optimal either, since it has the adverse effect of preventing the human operator from exercising her judgment effort. We characterize the liability sharing rule that achieves the highest possible quality level of the algorithm, while ensuring that the human operator exercises a judgment effort. We then show that, when it can be used, a negligence rule generally achieves the first best optimum. To conclude, we discuss the pros and cons of each type of liability rule.

Keywords: liability rules, decision-making, artificial intelligence, cognitive bias, judgment, prediction

JEL classification: K4.

*Université Paris Panthéon Assas, CRED EA 7321, 75005 Paris, France. e-mail: marie.obidzinski@u-paris2.fr

†Université Bourgogne Franche-Comté, CRESE EA 3190, Besançon, France. e-mail:yves.oytana@univ-fcomte.fr

The authors thank the participants at the 70th Congress of the French Economic Association annual conference, the German Law and Economics annual conference in Nancy, and the seminar organized by the ZiF Research Group on Economic and Legal Challenges in the Advent of Smart Products.

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-making support in a wide variety of situations, while *performative* algorithms are "able to accomplish independent actions by gathering information, decide and execute" (Jussupow et al., 2020). These two types of algorithms (advisory versus performative) are distinguished from each other by their degree of autonomy. For instance, in aviation, pilots have been transformed from operators to supervisors as a consequence of increased reliance on performative algorithms. We can expect to observe the same trend for drivers in automated cars in the coming years. Conversely, advisory algorithms leave the ultimate decision to the user. The focus of this paper will be on advisory algorithms, which are used in a wide range of contexts. Physicians use them to better interpret X-ray pictures, to better predict the occurrence and/or the evolution of a disease, and to better anticipate emerging infectious disease epidemics.[1] Advisory algorithms are used by judges to help them assess the risk of repeat offending by criminal defendants (*e.g.* the controversial COMPAS algorithm, for "Correctional Offender Management Profiling for Alternative Sanctions").[2] Banking institutions also use predictions given by credit scoring models to help them reduce the risk of debt default or to prevent fraud.

These algorithms provide for predictions to human operators, who then decide how to proceed with the additional information. Advisory algorithms can display varying degrees of sophistication, but they basically follow one of the two possible approaches: they may reproduce the cognitive mechanisms of a human expert (symbolic or rule-based approach), or they may look for regularities in data that are used to extract knowledge, without a preestablished model (digital or machine learning based approach).[3] With the latter approach, the algorithm is trained to generate a model. Deep learning algorithms follow that approach by learning through "trial and error". This learning often comes at the cost of great complexity to understand what the determinants of the algorithm's prediction are, and how these determinants interact with each other. Despite this, when using such advisory algorithms, human users are still in charge of the decision-making: following the terminology of Aghion

[1] See for instance Chopard and Musy (2022) who study the market for AI systems in health care.

[2] COMPAS has been criticized on the ground that the algorithm is allegedly biased against black defendants (see the propublica article by Larson, Mattu, Kirchner, and Angwin, 2016).

[3] Source: Inserm, 2018. A rule-based approach is by essence deterministic, contrary to a machine learning approach.

and Tirole (1997), the formal authority is held by the human operator.[4]

As AI algorithms often outperform humans in a wide variety of applications, large benefits are expected from their use. For instance, AI algorithms perform better than humans in the context of pretrial release decisions (Kleinberg et al., 2018) or in the context of medical imaging analyses like computed tomography (Cheng et al., 2016). By reducing the cost of providing high quality 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 support of an advisory algorithm, human decisions are prone to mistakes. This is especially true when the human operator fails to properly consider the reliability of the algorithm's prediction. Indeed, as is well recognized in the computer science literature, humans may be over-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 denoted as "automation bias", notably by Cummings (2017) and Mosier and Skitka (2018).[6] In our paper, the control problem or "misuse" (denoted *overestimation bias* in the following) is characterized by an overestimation of the probability that the prediction of the algorithm is correct.[7] The bias may lead the human operator to make an inappropriate use of the algorithm, and the AI manufacturer to invest insufficiently in its quality. Both of these choices affect the risk that a wrong decision will be taken, which may produce an external harm.[8]

One possible way to reduce the risk of external harm is to implement accountability mechanisms to share the burden of a poor outcome of human-AI collaborative decision-making.

[4]See Athey et al. (2020) for an analysis of a situation in which a principal has to choose to allocate the decision-making authority either to an IA or to a human agent.

[5]More specifically, misuse is defined by Parasuraman and Riley (1997) as over-reliance on automation.

[6]Although we do not consider this issue, note that after observing mistakes, people may also become prone to algorithm aversion (Zhang et al., 2020, Jussupow et al., 2020).

[7]The explainable AI approach tends to address the issue of over-reliance by providing insights into how the algorithm makes its prediction. However, this approach has not been very successful in achieving that goal (Buçinca et al., 2021).

[8]To illustrate, an inappropriate medical treatment may harm a third party, namely the patient.

The design of well-adapted legislation can provide the incentives to produce high quality algorithms while ensuring an adequate use of the predictions by human operators, thus reducing the risk of damage. The topic of AI regulation is on the agenda of many countries.[9]

Research questions. Our contribution addresses the issue of the optimal liability rule (sharing of liability and negligence rule) when an AI manufacturer develops an algorithm that will be used by a human operator. The quality 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, if so, whether to make a non-observable cognitive effort (which we will call *judgment effort* in the following). Finally, a decision relying 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 (observed by means of the judgment effort) is made. Absent any cognitive bias, a strict liability of the human operator makes her use the algorithm in an appropriate way and, in addition, incentivizes the AI manufacturer to make the socially optimal investment in the algorithm's quality (since he fully internalizes the expected liability cost through the price). On that premise, we ask the following questions: Might the overestimation bias justify a sharing of liability between a human operator and an AI manufacturer? Would a negligence rule pertaining to the AI manufacturer be optimal?

Assumptions and main results. In our model, we assume that the user of an algorithm suffers from an overestimation bias that results in the misperception of the risk of a wrongful prediction. More specifically, the user tends to overestimate the accuracy of the algorithm. As explained by Miceli and Segerson (2021) and following Zeiler (2019), that bias constitutes a "psychological mistake"[10] that affects the actual decisions made by the human operator,

[9] In Europe, as noted by Ebers (2021), "there is currently no specific legislation on civil liability for damage caused by AI either at European level or in any national jurisdictions." Legislation could take the form of regulation (with certification), defective product liability, and specific AI tort law. More precisely, in the European Union, "product liability has been fully harmonized in all Member States through the Product Liability Directive (PLD) 85/374/EEC which establishes a system of strict liability – that is, liability without fault for producers when a defective product causes physical or material damage to the injured person", as explained by Ebers (2021). However, there are numerous limits to the application of strict liability (such as the scope of the directive and the burden of proof). Liability for AI systems can arise also from national liability systems, especially tort law. In this domain, there is a wide range of approaches in Member States. The EU is working on the future legal framework (see for instance the Artificial Intelligence Act, and the ongoing revision of the Product Liability Directive).

[10] As explained by these authors, another type of "bias" comes from non-standard preferences. In contrast to a misperception bias, a bias resulting from non-standard preferences does not imply that the individual makes a mistake and thus should not necessarily be corrected.

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." This is why we consider the public authority in charge of choosing the liability rule as being a "paternalistic" one rather than a "populist" one (Salanié and Treich, 2009), meaning that this public authority computes welfare using the true probability of a wrongful prediction, as opposed to the probability perceived by the human operator who suffers from an overestimation bias.

Following Agrawal et al. (2019b), we assume that the prediction of the algorithm and the judgment effort made by the human operator cover two different dimensions. While the information given by the prediction relates to the actual state of the world (*e.g.* whether or not a patient has cancer), the judgment effort relates to the payoffs in each possible state of the world (*e.g.* whether or not the patient will benefit from intensive and expensive treatment in this specific case).

Our main results are the following. If the human operator does not suffer from an overestimation bias, a strict liability of the human operator is an optimal rule. Indeed, the expected cost is fully internalized in the price of the algorithm, resulting in the socially optimal investment in the quality of the algorithm, a proper adoption of the algorithm and an optimal level of judgment effort. In contrast, we show that if the human operator suffers from an overestimation bias, a strict liability of the human operator cannot be optimal, because the AI manufacturer will take advantage of the misperception of the operator, resulting in too low a level of quality. Nevertheless, strict liability of the AI manufacturer may not be optimal either, since in this case the human operator will never exert her judgment effort. In this case, we find that there is an optimal sharing of liability (which we characterize) such that the human operator will exert her judgment effort, while reducing the gap between the socially optimal level of quality of the algorithm and the one chosen by the AI manufacturer. Finally, we show that a negligence rule will generally incentivize the AI manufacturer to choose the socially optimal level of quality, and we discuss the conditions under which such a negligence rule can be implemented.

The rest of the paper is structured as follows. Sections 2 introduces the related literature. Section 3 presents the model setup. Section 4 solves the model for the optimal liability

sharing rule and shows that a negligence rule generally achieves the first best optimum. Finally, section 5 concludes.

2 Related Literature

Our paper is at the crossroads of several branches of the literature. First, there is a specific literature on AI and decision-making. Second, the paper is related to the law and economics literature on product liability and consumer biases. Third, some authors have investigated more specifically optimal product liability when smart products (and more specifically self-driving cars) are involved in accidents.

AI and decision-making. 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 exert 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, by assuming that AI predictions complement human judgment, in order to study the efficiency of liability rules in this context, knowing that as a risky decision may lead to external damage, and both judgment effort and AI quality may affect the decision that is made and thus the occurrence of harm.

Further, some literature on algorithm regulation has focused on algorithm bias (see among others Abrardi et al., 2021, Rambachan et al., 2020, Liang et al., 2021). Indeed, a well-known problem with advisory algorithms is that they may produce errors which will bias their prediction, thus affecting certain groups of people more, *e.g.* because the algorithm reproduces biases that already exist in a low quality training dataset. Biases are thus a significant fairness concern for regulators. However, in our paper, we do not focus on algorithm bias, but rather on the human operator's cognitive bias. This bias will affect how she will interpret the prediction given by the algorithm, while the prediction itself is not biased. More specifically, we assume that the human operator suffers from an overestimation bias, with the consequence that she tends to be over-reliant on the algorithm prediction.

Product liability and consumer biases. Our contribution is related to the literature on product liability.[11] Hay and Spier (2005) show that it is optimal that the consumer,

[11] Daughety and Reinganum (2013) and Geistfeld (2009) provide surveys of the product liability literature. See also the seminal paper by Landes and Posner (1985). In the same vein, our contribution is linked to

when fully solvent, bears the full liability of the external harm. However, when consumers are insolvent, a "residual-manufacturer liability" in which 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, but our contributions differ on the rationale for this result. 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 quality. A sub-part of the literature on product liability has considered biased consumers.[12] The closest to our paper is the article of Friehe et al. (2020). Friehe et al. (2020) compare liability rules when consumers are present-biased. We share their results that consumer bias may provide a rationale for sharing liability between the consumer and the (monopolistic) manufacturer, and that a negligence rule may yield further efficiency gains. However, we differ in the specific bias we consider and in the nature of the choice made by consumers (consumer care *versus* judgment effort, the latter being a non observable cognitive effort).[13] In a broader presentation of the role of bias in economic models of law, Miceli and Segerson (2021) recently look at the case where consumers misperceive their risk of damage both in 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 a sharing of liability may be optimal when consumers underestimate the risk, because it offsets (in part) the monopoly distortion on quantities. By contrast, we do not consider how liability rules may result in under or overproduction since, in our model, users buy at most one prediction and are not heterogeneous in their willingness to pay for it. Rather, what drives our result that strict manufacturer liability may not be optimal is that it induces a crowding out effect on the user's judgment effort.

Autonomous vehicles (AV) and liability rules. There is an emerging pool of literature on the liability rules that should apply to autonomous vehicles (Shavell, 2020; Talley, 2019; De Chiara et al., 2021; Dawid and Muehlheusser, 2022; Guerra et al., 2021a,b). These

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 mainly concerned with the dilution of liability and the risk of suboptimal care.

[12] On consumer misperceptions, see the seminal papers of Spence (1977) and Polinsky and Rogerson (1983).

[13] See also Baniak and Grajzl (2017) who investigate the consequences of possible customer misperceptions about future usage, denoted *projection bias*.

contributions differ regarding (1) whether or not the probability of accident can be affected by the car'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.[14] Our framework is different, as we consider that the victim is passive and has no impact on the probability of the accident occurring. Moreover, we do not deal with performative algorithms such as autonomous cars, but with advisory algorithms. In spite of that, we share some important results with these contributions. Like Shavell (2020) and Talley (2019), we find that a strict liability regime is not well adapted. Moreover, similarly 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 quality. However, in our framework, increasing the share of liability borne by the manufacturer also leads to an under provision of effort by the human operator, as she has no incentive to exert it when she does not bear the external harm. The major difference with our contribution is that, to our knowledge, the emerging literature on AV has not yet taken into account the fact that human users may be prone to cognitive biases when they interact with machines.

3 Model Presentation

We build a model based on Agrawal et al. (2019a) (inspired by Bolton and Faure-Grimaud, 2009) in order to derive the optimal liability regime chosen by a policymaker (A) when the use of an algorithm may cause an external harm. The purpose of the liability regime is to apportion this harm 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 may lead the user of the algorithm to be overly confident in the prediction made by the AI.

Initially, A chooses a liability regime in order to maximize the expected welfare.[15] We restrict our attention to the following liability regimes: (i) a *shared liability rule* in which damages are apportioned between the AI manufacturer and the human operator (these possible sharing

[14]For instance, Shavell (2020) considers a model in which all vehicles are autonomous and suggests using a new form of liability in which damages are paid to the state. Talley (2019) and De Chiara et al. (2021) develop a model where 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. car market, study how liability rules may impact the emergence and the development of AV.

[15]Contrary to Talley (2019), Shavell (2020), De Chiara et al. (2021), Guerra et al. (2021a), and Guerra et al. (2021b), we consider a framework where the victim is a passive third party, that is, they have no impact on the probability of harm.

rules include the strict liability of M or H) and (ii) a *negligence rule* in which the standard specifies a minimum level of quality of the algorithm.[16] Then, M chooses the level of quality of the algorithm (q) and the price (p) at which it is sold to H. H chooses whether to buy the algorithm. If she does so, the algorithm gives her a prediction (which may be wrong). We model this prediction as a noisy signal about the actual state of the world. After observing that prediction, she chooses whether or not to exert a judgment effort which allows her to obtain information on the payoff that may be expected in each possible state of the world. Finally, a decision is made that may cause an external harm, depending on the realized state of the world.

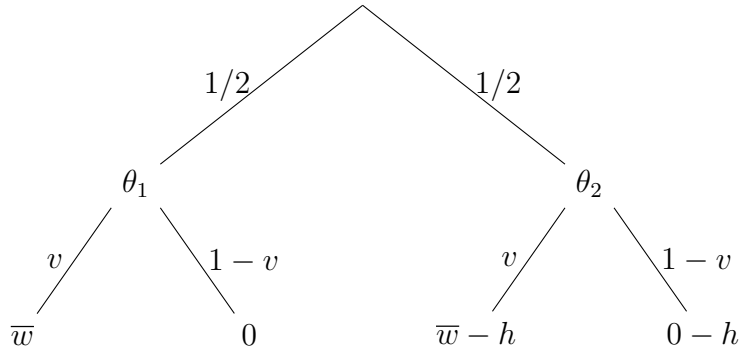
Because we are mainly interested in M's investment decision and H's decision whether or not to exercise a costly judgment, we assume that the decision made in the final stage of the game is a function of the algorithm's prediction and the information revealed when H exerts a judgment effort. Following Agrawal et al. (2018), that decision is denoted $y \in \{S, R\}$. If a safe decision $y = S$ is made, H's gain from the decision is normalized to 0. In that case, there is no risk of external harm h. If a risky decision $y = R$ is made, H's gain is given by $w \in \{\overline{w}, 0\}$, with $\overline{w} > 0$. H's gain w from a risky decision is equal to $\overline{w}$ (resp. 0) with probability v (resp. $1 - v$). However, a risky decision $y = R$ may cause an external harm h. The occurrence of that harm depends on the state of the world, which is denoted $\theta \in \{\theta_1, \theta_2\}$. We assume $\Pr(\theta = \theta_1) = \Pr(\theta = \theta_2) = 1/2$.[17] In state $\theta = \theta_1$, there is no external harm. The external harm h is realized only in the state of the world $\theta = \theta_2$. We assume the external harm h is higher than the expected gain vw from a risky decision (absent any judgment effort): $v\overline{w} - h < 0$.

Figure 1 summarizes the payoffs from a risky decision when no information on θ and w is available.

[16] The standard cannot be based on the judgment effort of the human operator. Indeed, the judgment effort is a cognitive effort which is by definition not observable nor verifiable. In Bolton and Faure-Grimaud (2009), judgment is depicted as a "thought experiment".

[17] For simplicity, following Agrawal et al. (2019b), we assume that each state of the world is equally likely to occur.

Figure 1: Payoff from a risky decision



If a shared liability rule is applied, the fraction of the external harm for which H (resp. M) is liable is α (resp. $1-\alpha$). Thus, if $\alpha = 0$, the liability lies entirely with M (strict liability of M), while it lies entirely with H if $\alpha = 1$ (strict liability of H). The damage sharing rule $\alpha \in [0,1]$ is chosen by A in order to maximize the expected welfare, which is computed by considering the objective probability that the algorithm divulges the true state of the world, rather than the probability perceived by H (which may be distorted upward due to an overestimation bias). If a negligence rule is applied, we assume the standard is equal to the level of quality of the algorithm that maximizes the expected welfare. If M chooses a level of quality that is higher than or equal to the standard, then the liability rests entirely on H ($\alpha = 1$). If M chooses a level of quality below the standard, then the liability rests entirely on M ($\alpha = 0$).

To obtain a prediction from the algorithm, H has to pay a price p (chosen by M). If she does so, she observes a signal $s_a \in \{\theta_1, \theta_2\}$ about the state of the world. The prediction technology is such that $\Pr(s_a = \theta) = \phi(q) \in [1/2, 1]$ (*i.e.* H learns the true state with probability $\phi(q)$), with $q \in [0,1]$ the quality of the algorithm, $\phi(0) = 1/2$, and $\phi'(q) > 0$. This quality of the algorithm is chosen by M and costs him $c(q)$, with $c'(0) = 0$, $c'(q) > 0$ if $q > 0$, $\lim_{q \to 1} c'(q) = +\infty$ and $c''(q) > 0$ if $q > 0$. We assume that H may be prone to an overestimation bias, in which case the probability that the algorithm divulges the correct state of the world, as perceived by the human operator, is estimated to be $\tilde{\phi}(q;\epsilon) > \phi(q)$ if $\epsilon > 0$, with $\epsilon \in [0,1]$ the magnitude of the overestimation bias. The higher the bias, the higher the subjective probability that the algorithm divulges the correct probability ($\frac{\partial \tilde{\phi}}{\partial \epsilon}(q;\epsilon) > 0$).[18] An increase in the level of quality of the algorithm (q) has a positive effect

[18]For instance, H's perceived probability of a correct prediction may be specified as $\tilde{\phi}(q;\epsilon) = (1-\epsilon)\phi(q) +$

on both the objective probability that the algorithm divulges the correct state of the world ($\phi'(q) > 0$) and on the probability perceived by H ($\tilde{\phi}'(q) > 0$), although the former effect is larger than the latter ($\phi'(q) > \tilde{\phi}'(q;\epsilon)$ if $\epsilon > 0$),[19] and that the magnitude of these increases declines with the level of quality ($\phi''(q) < 0$ and $\tilde{\phi}''(q;\epsilon) < 0$). Moreover, we assume that $\partial^2\tilde{\phi}/(\partial q \partial\epsilon)(q;\epsilon) < 0$.

After observing the prediction, H may exert a judgment effort at cost $k > 0$.[20] That effort allows her to learn about the payoff of a risky decision in each possible state of the world. More specifically, since there are two possible states of the world, the judgment effort allows H to observe one of four possible ordered pairs of payoffs from a risky decision: $(\overline{w}, \overline{w})$, $(\overline{w}, 0)$, $(0, \overline{w})$, $(0, 0)$. In each of these four pairs, the first (resp. second) object is the gross payoff ("gross" because it does not include a potential liability payment) obtained by H following a risky decision $y = R$ if $\theta = \theta_1$ (resp. $\theta = \theta_2$). In the following analysis, we will consider that a favorable prediction $s_a = \theta_1$ is a necessary condition for a risky decision to be made. A risky decision is then made, unless H exerts a judgment that reveals $(0, \overline{w})$ or $(0, 0)$, in which cases the safe decision $y = S$ is made.[21]

4 Analysis

We explore the choices under different liability schemes made by the two actors of the decision-making process, that is the AI manufacturer (M) and the human operator (H), before turning to the choice of the optimal liability scheme made by the policymaker. We

$\epsilon\overline{\phi}$, with $\overline{\phi} = \lim_{q\to 1}\phi(q)$.

[19] A possible interpretation is that due to the overestimation bias, H overestimates the level of quality of the algorithm, but is aware that there are some decreasing marginal returns from increasing the level of quality (q) on the probability of obtaining a correct prediction. Consequently, she will anticipate a smaller marginal effect of q on her perceived probability $\tilde{\phi}(q;\epsilon)$ than what she would have anticipated by considering the objectively correct probability $\phi(q)$. In other words, H overestimates the quality of the algorithm, but not its marginal effect (or at least the sign of that marginal effect) on the prediction reliability.

[20] We implicitly assume in the model that the cost of the judgment effort (k) is known and is the same for all users. If users were heterogeneous with respect to this cost and if no price discrimination were possible, the AI manufacturer would face the standard trade-off between lowering the price to sell more units and increasing it to get a higher markup per unit. In this scenario, only a fraction of the potential users would buy the algorithm, and the determination of the optimal liability sharing rule would be more complex. More specifically, we could expect that by decreasing (resp. increasing) the share of liability borne by H, fewer (resp. more) users would acquire the algorithm, generating excessive or insufficient algorithm adoption.

[21] Somehow, judgment is *optional*. Alternatively, we could have considered that judgment is mandatory before taking any risky decision. However, the aim of the paper is to address the issue of misuse of algorithms' predictions, where the user is over-reliant on algorithms' predictions. Therefore, the optional judgment hypothesis is more suitable.

assume that liability can be shared *ex ante*, where H faces the share of liability α and M the remaining part $1-\alpha$ (Hay and Spier, 2005; Friehe et al., 2020).[22] In a first subsection, we derive the optimal sharing rule. In a second subsection, we show that a negligence rule allows the first best optimum to be restored.

4.1 Sharing of liability between the human user and the manufacturer

With sharing of liability, two possible types of equilibria may arise. In the first one (denoted $E0$) in the following, H does not exert her judgment effort, and conversely in the second one (denoted $E1$). In what follows, we first study the judgment effort choice (section 4.1.1), before turning our attention to each one of these possible equilibria (sections 4.1.2 and 4.1.3) and then comparing these cases to find the the optimal liability sharing rule (section 4.1.4).

4.1.1 The human operator's judgment effort

Recall that a "favorable" algorithm prediction ($s_a = \theta_1$) is a necessary condition for a risky decision $y = R$ to be made. Otherwise, a safe decision $y = S$ is always made, which brings zero expected utility to H.

After receiving a favorable prediction ($s_a = \theta_1$), H will exert a judgment effort only if it brings her a higher expected utility. Upon observing $s_a = \theta_1$, H's perceived expected utility when she chooses *not* to exert a judgment effort is:

$$\begin{aligned} u_0(q,\alpha;\epsilon) &= \tilde{\phi}(q;\epsilon)v\overline{w} + (1-\tilde{\phi}(q;\epsilon))(v(\overline{w}-\alpha h) + (1-v)(-\alpha h)) \\ &= v\overline{w} - (1-\tilde{\phi}(q;\epsilon))\alpha h \end{aligned} \tag{1}$$

The first (resp. the second) term of (1) is H's perceived expected utility from a risky decision in the event the algorithm makes an accurate prediction $s_a = \theta_1 = \theta$ (resp. in the event the algorithm makes an inaccurate prediction $s_a = \theta_1 \neq \theta$). H bears some liability cost only if $\alpha > 0$ (M is not strictly liable) and if the algorithm's prediction is incorrect, which arises with a perceived probability of $1-\tilde{\phi}(q;\epsilon)$.

Let us now consider H's expected utility when she chooses to exert a judgment effort, given

[22] In Hay and Spier (2005), $1-\alpha$ is denoted "residual" manufacturer liability, in contrast to "consumer-only" liability (*i.e.* $\alpha = 1$).

the observation of a prediction $s_a = \theta_1$. Recall that a risky decision $y = R$ is then made if H's judgment effort reveals $(\overline{w}, \overline{w})$ or $(\overline{w}, 0)$. If H's judgment effort reveals $(0, \overline{w})$ or $(0, 0)$ the safe decision $y = S$ is made. Therefore, H's expected utility if she decides to exert a judgment effort is given by:

$$u_1(q, \alpha; \epsilon) = v^2 \left[\overline{w} - (1 - \tilde{\phi}(q; \epsilon))\alpha h\right] + v(1 - v) \left[\tilde{\phi}(q; \epsilon)\overline{w} - (1 - \tilde{\phi}(q; \epsilon))\alpha h\right] - k \qquad (2)$$

The first (resp. the second) term of (2) is H's perceived expected utility if her judgment effort reveals $(\overline{w}, \overline{w})$ (resp. $(\overline{w}, 0)$). The third term is the cost of the judgment effort. With a judgment effort, the conditions that have to be fulfilled to take the risky decision are more demanding than without the judgment effort, as the payoff revealed by the judgment effort in state $\theta = \theta_1$ should be $\overline{w}$. Therefore, the risk of external harm (which cannot happen under the safe decision) is reduced.

Note that, regardless of the judgment effort, if H suffers from a bias ($\epsilon > 0$), she underestimates the risk of the algorithm's prediction being incorrect. More specifically, an increase in the overestimation bias implies that, when H observes $s_a = \theta_1$, her perceived expected utility is higher than her objective expected utility, and the gap between the two increases, as $\frac{\partial \tilde{\phi}}{\partial \epsilon}(q; \epsilon) > 0$.

H prefers *not* to make the judgment effort if:

$$u_1(q, \alpha; \epsilon) < u_0(q, \alpha; \epsilon) \Leftrightarrow k > (1 - \tilde{\phi}(q; \epsilon))(1 - v)(\alpha h - v\overline{w}) \equiv k_H(q, \alpha; \epsilon) \qquad (3)$$

For a given level of quality of the algorithm (q), the threshold of the judgment cost above which H prefers *not* to make the judgment is increasing with her share of liability (α), thus increasing the set of the judgment costs $k \in [0; k_H(q, \alpha; \epsilon)]$ for which H exerts a judgment. An increase in the quality of the algorithm (q) or in the magnitude of the overestimation bias (ϵ) has an ambiguous effect on that threshold. More specifically, if the share of liability lying with H is large (high α), then an increase in the overestimation bias or in the quality of the algorithm decreases the set of the judgment costs for which H exerts a judgment (and conversely if α is low).

Let us consider now the particular case where the manufacturer M bears the full liability in the event of a damage ($\alpha = 0$).

Lemma 1. *H never makes a judgment effort when the manufacturer is strictly liable ($\alpha = 0$).*

Proof. The proof results from the fact that $k_H(q, 0; \epsilon) < 0$. □

The intuition is that if M is fully liable for any eventual harm, H does not internalize the expected harm associated with a risky decision. Thus, avoiding the judgment effort is less costly for H and allows her to promote the risky decision, which brings her a positive expected gain.

4.1.2 E0: the human operator does *not* exert her judgment effort

Let us consider the case where H does not exert any judgment effort: we assume that (3) is satisfied (we will discuss at the end of this subsection to what extent this is indeed the case).

H's choice to buy the algorithm. If H does not acquire the algorithm, the prediction $s_a = \theta_1$ is never observed. As a consequence, H obtains an expected utility of 0. Conversely, if H acquires the algorithm (knowing that she will not exert a judgment effort even if the prediction is $s_a = \theta_1$), her expected utility is:

$$U_0(q, \alpha, p; \epsilon) = \frac{1}{2} u_0(q, \alpha; \epsilon) - p \tag{4}$$

The probability of 1/2 in the expression above is the probability that the prediction is $s_a = \theta_1$.[23]

H's willingness to pay for the algorithm's prediction is:

$$U_0(q, \alpha, p; \epsilon) = 0 \Leftrightarrow p = \frac{1}{2} v\overline{w} - \frac{1}{2}(1 - \tilde{\phi}(q; \epsilon))\alpha h \equiv p_0(q, \alpha; \epsilon) \tag{5}$$

Unsurprisingly, the lower H's share of liability and the higher her overestimation bias, the higher is her willingness to pay for the algorithm. Note also that this willingness to pay for the prediction is higher than both the social value of the prediction and its true value for the

[23] Indeed, if H pays to observe a prediction, her perceived probability of observing a prediction $s_a = \theta_1$ (which is biased due to the overestimation bias) is:

$$\begin{aligned} \Pr(s_a = \theta_1) &= \Pr(\theta = \theta_1) \times \Pr(s_a = \theta_1 | \theta = \theta_1) + \Pr(\theta = \theta_2) \times \Pr(s_a = \theta_1 | \theta = \theta_2) \\ &= \frac{1}{2}\tilde{\phi}(q; \epsilon) + \frac{1}{2}(1 - \tilde{\phi}(q; \epsilon)) \\ &= \frac{1}{2} \end{aligned}$$

user (which are respectively $p_0(q, 1; 0)$ and $p_0(q, \alpha; 0)$). As a consequence, if the manufacturer sells the algorithm at price $p_0(q, \alpha; \epsilon)$, the *ex post* expected utility of H is negative: the AI manufacturer's pricing policy is *consumer exploitative* (Bienenstock, 2016), in the sense that the AI manufacturer takes advantage of the user's overestimation bias to increase his profit.[24]

M's choice of price and whether to distribute the algorithm. We assume that M is a monopolist.[25] He has the possibility of selling his algorithm at a maximum price of $p_0(q, \alpha; \epsilon)$. However, it is possible that this price does not allow him to obtain a positive expected profit. In this case, he may choose to not distribute the algorithm (or equivalently to set a price $p > p_0(q, \alpha; \epsilon)$). If the algorithm is not distributed, then there is no point in M developing the algorithm: he chooses a level of quality $q = 0$ and obtains an expected profit equal to 0.

Let us assume for now that M does have an interest in developing and distributing the algorithm (we will discuss at the end of this subsection to what extent this is indeed the case).

M's choice as to the level of quality when the algorithm is distributed. The expected profit of M is the price he obtains from the sale of the algorithm ($p_0(q, \alpha; \epsilon)$), minus the expected liability he faces in case of external harm, minus the cost of his investment in the quality of the algorithm:

$$\pi_0(q, \alpha; \epsilon) = p_0(q, \alpha; \epsilon) - \frac{1}{2}(1 - \phi(q))(1 - \alpha)h - c(q) \tag{6}$$

If the algorithm development and distribution brings a positive expected profit and if M anticipates that H will not exert a judgment effort even if $s_a = \theta_1$, he chooses a level of quality $q^*_{M,0}(\alpha; \epsilon) = \arg\max_q \pi_0(q, \alpha; \epsilon)$.[26]

[24] We do not consider the possibility that the AI manufacturer may be willing to "debias" or educate the users of their algorithm.

[25] By imposing a zero-profit condition on the AI manufacturer, it is possible to show that H's decision to exercise a judgment effort and M's choice of quality remain identical. The main difference is the price chosen by M. However, from society's perspective, this price is a mere monetary transfer and therefore is irrelevant. Consequently, whether the market in which the algorithm is sold is monopolistic or is perfectly competitive has very little impact on our results.

[26] The FOC is:

$$\frac{\partial \pi_0}{\partial q}(q, \alpha; \epsilon) = 0 \Leftrightarrow \frac{h}{2}\left(\frac{\partial \phi}{\partial q}(q) + \left(\frac{\partial \tilde{\phi}}{\partial q}(q; \epsilon) - \frac{\partial \phi}{\partial q}(q)\right)\alpha\right) = c'(q) \tag{7}$$

A's choice of the liability sharing rule. The expected social welfare is the sum of H's *ex post* utility $U_0(q, \alpha; 0)$ and M's expected profit $\pi_0(q, \alpha; \epsilon)$:

$$W_0(q) = \frac{1}{2}v\overline{w} - \frac{1}{2}(1 - \phi(q))h - c(q) \tag{8}$$

This expected social welfare is equal to M's expected profit when H does not suffer from an overestimation bias: $\pi_0(q, \alpha; 0) = W_0(q)$. The socially optimal level of quality is given by $q^*_{W,0} = \arg\max_q W_0(q)$.[27]

M's expected profit can be rewritten:

$$\pi_0(q, \alpha; \epsilon) = W_0(q) + \frac{1}{2}\left(\tilde{\phi}(q; \epsilon) - \phi(q)\right)\alpha h \tag{10}$$

Absent any overestimation bias ($\epsilon = 0$), M perfectly internalizes the expected harm whatever the liability sharing rule that applies (because in this case $\tilde{\phi}(q; \epsilon) = \phi(q)$), while he internalizes only a fraction of it if $\epsilon > 0$. Comparing the FOCs for $q^*_{M,0}(\alpha; \epsilon)$ and $q^*_{W,0}$, we have:

$$\frac{\partial \pi_0}{\partial q}(q, \alpha; \epsilon) - \frac{\partial W_0}{\partial q}(q) = \frac{1}{2}\left(\frac{\partial \tilde{\phi}}{\partial q}(q; \epsilon) - \frac{\partial \phi}{\partial q}(q)\right)\alpha h \tag{11}$$

Thus, if $\epsilon = 0$, the quality of the algorithm chosen by M is socially optimal. In contrast, if $\epsilon > 0$, the quality of the algorithm is socially optimal only if $\alpha = 0$ (strict liability of M).

Discussion of the conditions under which $E0$ may emerge. We have found that when (i) H is not willing to exert a judgment effort even when $s_a = \theta_1$ and (ii) M obtains a positive expected profit by developing and selling the algorithm at a price equal to H's willingness to pay, then the socially optimal sharing rule chosen by A is $\alpha = 0$, so that the quality level chosen by M is $q^*_{M,0}(0; \epsilon) = q^*_{W,0}$. Are conditions (i) and (ii) compatible with a strict liability of M ($\alpha = 0$)?

Regarding (i), assuming that $\alpha = 0$, we know from lemma 1 that H will have no interest in making a judgment effort.

Moreover, note that $\frac{\partial^2 \pi_0}{\partial q^2}(q, \alpha; \epsilon) < 0$, which implies that if M cannot choose a quality $q^*_{M,0}(\alpha; \epsilon)$, then he chooses the closest possible level of quality.

[27] The FOC is:

$$\frac{\partial W_0}{\partial q}(q) = 0 \Leftrightarrow \frac{1}{2}\frac{\partial \phi}{\partial q}(q)h = c'(q) \tag{9}$$

Regarding (ii), assuming that $\alpha = 0$, we know from (10) that $\pi_0(q, 0; \epsilon) = W_0(q)$. Thus, $\pi_0(q^*_{W,0}, 0; \epsilon) = W_0(q^*_{W,0})$ and provided that the expected social welfare is positive at the first best (*i.e.* $W_0(q^*_{W,0}) \geq 0$), then if M is strictly liable, he will choose a quality of the algorithm $q = q^*_{M,0}(0; \epsilon) = q^*_{W,0}$ and sell the algorithm at price $p = p_0(q^*_{W,0}, 0; \epsilon)$. This solution coincides with the first best optimum. Conversely, if $W_0(q^*_{W,0}) < 0$, then the quality of the algorithm chosen by M is $q = 0$ (the algorithm is not developed) and the price of the algorithm is $p > p_0(0, 0; \epsilon)$ (the algorithm is not distributed). M gives up the development and the distribution of the algorithm in order to obtain a null expected profit, which also coincides with the first best optimum. Lemma 2 resumes these results.

Lemma 2. *Given $\alpha = 0$ (strict liability of M), the quality of the algorithm chosen by M is $q = q^*_{M,0}(0; \epsilon) = q^*_{W,0}$ and the algorithm is sold at price $p = p_0(q^*_{W,0}, 0; \epsilon)$ if $W_0(q^*_{W,0}) \geq 0$. If $W_0(q^*_{W,0}) < 0$, the algorithm is not developed ($q = 0$) and is not distributed ($p > p_0(0, 0; \epsilon)$).*

Note that if $\epsilon = 0$, then the expected profit of M is equivalent to the expected social welfare regardless of the liability sharing rule (*i.e.* $\pi_0(q, \alpha; 0) = W_0(q)$). This contrasts with the results we get if $\epsilon > 0$, in which case M's expected profit is equivalent to the expected social welfare only if $\alpha = 0$ (strict liability of M). This means that if $\epsilon = 0$ and independently of the liability sharing rule, the quality chosen by M is always $q^*_{W,0}$ if for that level of quality H does not exert a judgment effort and M's expected profit is positive. Thus, if H does not suffer from an overestimation bias ($\epsilon = 0$), the strict liability of H may be socially optimal, which is never the case if $\epsilon > 0$.

4.1.3 E1: the human operator exerts her judgment effort

Let us next consider the case where H exerts a judgment effort: we assume that (3) is not satisfied (*i.e.* $k \leq k_H(q, \alpha; \epsilon)$) and will again discuss at the end of this subsection to what extent this is indeed the case.

H's choice to buy the algorithm. As in case $E0$, if H does not acquire the algorithm, the prediction $s_a = \theta_1$ is never observed and H obtains an expected utility of 0. Conversely, if H acquires the algorithm, she exerts a judgment effort if $s_a = \theta_1$ and thus her expected utility is:

$$U_1(q, \alpha, p; \epsilon) = \frac{1}{2} u_1(q, \alpha; \epsilon) - p \tag{12}$$

H's willingness to pay for the algorithm's prediction is:

$$\begin{aligned} &U_1(q,\alpha,p;\epsilon)=0 \\ &\qquad \Leftrightarrow p=\frac{1}{2}\left(v\left(\tilde{\phi}(q;\epsilon)+\left(1-\tilde{\phi}(q;\epsilon)\right)v\right)\overline{w}-k\right)-\frac{1}{2}\left(1-\tilde{\phi}(q;\epsilon)\right)v\alpha h \equiv p_1(q,\alpha;\epsilon) \end{aligned} \tag{13}$$

As in $E0$, the lower H's share of liability and the higher her overestimation bias, the higher is her willingness to pay for the algorithm.

M's choice of price and whether to distribute the algorithm. By similar reasoning to that of $E0$, M will sell the algorithm at price $p_1(q,\alpha;\epsilon)$ if this allows him to obtain a positive expected profit. Otherwise, he will not develop the algorithm. We assume for now that M benefits from developing and distributing the algorithm (we discuss the relevance of this assumption at the end of this subsection).

M's choice as to the level of quality when the algorithm is distributed. The expected profit of M is:

$$\pi_1(q,\alpha;\epsilon)=p_1(q,\alpha;\epsilon)-\frac{1}{2}(1-\phi(q))v(1-\alpha)h-c(q) \tag{14}$$

When comparing the expression of M's profit in the absence of any judgment effort, defined in (6), and M's profit with judgment effort, defined in (14), we see that the expected liability faced by M in $E1$ is scaled down by a factor v when compared to $E0$. This is because, in $E1$, H exerts a judgment effort which may reveal a payoff $w=0$ in state $\theta=\theta_1$, in which case the safe decision is made and the external harm that could have resulted from a wrongful prediction is avoided.

If M chooses to develop and distribute the algorithm, the quality of the algorithm is $q^*_{M,1}(\alpha;\epsilon)=\arg\max_q \pi_1(q,\alpha;\epsilon)$.[28]

A's choice of the liability sharing rule. The expected social welfare is:

$$W_1(q)=\frac{1}{2}\left(v(\phi(q)+(1-\phi(q))v)\overline{w}-k\right)-\frac{1}{2}(1-\phi(q))vh-c(q) \tag{16}$$

[28]If M anticipates that H will exert a judgment rather than not, what is the effect on quality of the algorithm? From (20), we know that $q^*_{W,0}>q^*_{W,1}$. Moreover, we know that $q^*_{M,0}(0;\epsilon)=q^*_{W,0}$ (lemma 2) and

The socially optimal level of quality is $q^*_{W,1} = \arg\max_q W_1(q)$.

M's profit can be rewritten:

$$\pi_1(q, \alpha; \epsilon) = W_1(q) + \frac{v}{2}\left(\tilde{\phi}(q; \epsilon) - \phi(q)\right)((1 - v)\overline{w} + \alpha h) \tag{17}$$

As in case $E0$, this expected social welfare is equal to M's expected profit when H does not suffer from an overestimation bias: $\pi_1(q, \alpha; 0) = W_1(q)$.

Lemma 3. *Assume that H always makes a judgment effort and that M's expected profit from developing and distributing the algorithm is positive. (i) If $\epsilon = 0$ (H does not have an overestimation bias), the quality of the algorithm chosen by M is socially optimal independently of the liability sharing rule (i.e. $q^*_{M,1}(\alpha; 0) = q^*_{W,1}$ $\forall \alpha \in [0, 1]$). (ii) If $\epsilon > 0$ (H has an overestimation bias), the quality of the algorithm chosen by M is less than the socially optimal level of quality (i.e. $q^*_{M,1}(\alpha; \epsilon) < q^*_{W,1}$ $\forall \alpha \in [0, 1]$) and is decreasing with the share of liability borne by H ($\frac{\partial q^*_{M,1}}{\partial \alpha}(\alpha; \epsilon) < 0$).*

Proof. The proof is in the appendix (section 6.1). □

In other words, if H suffers from an overestimation bias ($\epsilon > 0$), the marginal benefit for M of increasing the quality of the algorithm is less than the social marginal benefit: the quality chosen by M is thus insufficient. Moreover, this distortion between the quality chosen by M and the socially optimal quality increases when the share of liability borne by H increases. This implies that, if H has an overestimation bias, as long as the policymaker anticipates that H will make a judgment effort given M's best response $q = q^*_{M,1}(\alpha; \epsilon)$, it will be socially beneficial to increase M's share of liability (*i.e.* to decrease α). Conversely, when H does not

$q^*_{M,1}(0; \epsilon) \leq q^*_{W,1}$ (lemma 3). Together, this implies that $q^*_{M,0}(0; \epsilon) > q^*_{M,1}(0; \epsilon)$. Moreover, we have:

$$\begin{aligned}&\frac{\partial \pi_1}{\partial q}(q^*_{M,0}(\alpha; \epsilon), \alpha; \epsilon) - \frac{\partial \pi_0}{\partial q}(q^*_{M,0}(\alpha; \epsilon), \alpha; \epsilon) \\ &\qquad = -\frac{1}{2}(1 - v)\left(\left(\frac{\partial \tilde{\phi}}{\partial q}(q^*_{M,0}(\alpha; \epsilon); \epsilon)\alpha + \frac{\partial \phi}{\partial q}(q^*_{M,0}(\alpha; \epsilon))(1 - \alpha)\right)h - \frac{\partial \tilde{\phi}}{\partial q}(q^*_{M,0}(\alpha; \epsilon); \epsilon)v\overline{w}\right)\end{aligned} \tag{15}$$

Since the expected profit $\pi_1(q, \alpha; \epsilon)$ is concave with respect to the level of quality ($\frac{\partial^2 \pi_1}{\partial q^2}(q, \alpha; \epsilon) < 0$), the judgment effort has a negative impact on the quality of the algorithm if (15) is negative. Because, as shown above, $q^*_{M,0}(0; \epsilon) > q^*_{M,1}(0; \epsilon)$, (15) is negative for $\alpha = 0$. Moreover, we can show that (15) is decreasing with α. Thus, for all $\alpha \in [0, 1]$, (15) is negative: the manufacturer chooses a higher level of quality when he anticipates that the human operator will *not* exert her judgment.

have an overestimation bias ($\epsilon = 0$), the choice of M regarding the quality of the algorithm is always socially optimal.

Lemma 4. *Assume that H always makes a judgment effort and that M's expected profit from developing and distributing the algorithm is positive. (i) If $\epsilon = 0$ (H does not have an overestimation bias), the expected social welfare is independent of the sharing of liability between H and M (i.e. $\frac{\mathrm{d}W_1(q^*_{M,1}(\alpha;0))}{\mathrm{d}\alpha} = 0$). (ii) If $\epsilon > 0$ (H has an overestimation bias), the expected social welfare is strictly increasing with the share of liability borne by M (i.e. $\frac{\mathrm{d}W_1(q^*_{M,1}(\alpha;\epsilon))}{\mathrm{d}\alpha} < 0$).*

Proof. The proof is in the appendix (section 6.2). □

Lemmas 1 and 4 imply that there is a trade-off when choosing the liability sharing rule. On the one hand, we know from lemma 4 that if H suffers from an overestimation bias ($\epsilon > 0$), the choice of the liability sharing rule is not neutral, since by increasing the share of liability that falls on M, A obtains a socially beneficial improvement in the quality of the algorithm. On the other hand, we know from lemma 1 that decreasing too much the liability of H (for instance by choosing a strict liability of M, with $\alpha = 0$) may disincentivize H to make a judgment effort.

Finally, the liability sharing rule that maximizes the quality of the algorithm (assuming it exists) while ensuring that H will indeed exert a judgment effort, denoted $\alpha^*_{W,1}(\epsilon)$, is characterized by:

$$k = k_H(q^*_{M,1}(\alpha^*_{W,1}(\epsilon);\epsilon), \alpha^*_{W,1}(\epsilon);\epsilon) \tag{18}$$

The following lemma specifies under which condition such a rule exists.

Lemma 5. *There exists a threshold value for the cost of the judgment effort (k), denoted $k^*_1(\epsilon)$, such that if $k \leq k^*_1(\epsilon)$ (resp. $k > k^*_1(\epsilon)$), then $\alpha^*_{W,1}(\epsilon) \in (0,1]$ (resp. $\alpha^*_{W,1}(\epsilon) > 1$).*

Proof. The proof is in the appendix (section 6.3). □

Discussion of the conditions under which $E1$ may emerge. Lemma 5 implies that if $k > k^*_1(\epsilon)$, then there is no liability sharing rule $\alpha \in [0,1]$ such that, given the quality level chosen by M, H will exert a judgment effort. Conversely, if $k \leq k^*_1(\epsilon)$, lemma 5 guarantees that the sharing of liability $\alpha^*_{W,1}(\epsilon)$ is implementable by A and is such that H will exert a judgment effort. Given this liability sharing rule, M develops the algorithm with a level of

quality $q^*_{M,1}(\alpha^*_{W,1}(\epsilon);\epsilon)$ and sells it at price $p^*_1(q^*_{M,1}(\alpha^*_{W,1}(\epsilon);\epsilon),\alpha^*_{W,1}(\epsilon);\epsilon)$ if, as a result, he obtains a positive expected profit.

The following lemma ensures that if the expected social welfare is positive for a given quality level q, then M's expected profit will also be positive regardless of the liability sharing rule that applies.

Lemma 6. *(i) If $\epsilon = 0$ (H does not have an overestimation bias), $\pi_1(q,\alpha;\epsilon) = W_1(q)$ $\forall q$ and $\forall \alpha$. (ii) if $\epsilon > 0$ (H has an overestimation bias), $\pi_1(q,\alpha;\epsilon) > W_1(q)$.*

Proof. The proof is in the appendix (section 6.4). □

Thus, from lemma 6, if $W_1(q^*_{M,1}(\alpha^*_{W,1}(\epsilon);\epsilon)) \geq 0$, then $\pi_1(q^*_{M,1}(\alpha^*_{W,1}(\epsilon);\epsilon),\alpha^*_{W,1}(\epsilon);\epsilon) \geq 0$. We can thus formulate the following lemma.

Lemma 7. *Assume that $k \leq k^*_1(\epsilon)$ and $\alpha = \alpha^*_{W,1}(\epsilon)$. If $W_1(q^*_{M,1}(\alpha^*_{W,1}(\epsilon);\epsilon)) \geq 0$, the quality of the algorithm chosen by M is $q^*_{M,1}(\alpha^*_{W,1}(\epsilon);\epsilon)$ and the algorithm is sold at price $p^*_1(q^*_{M,1}(\alpha^*_{W,1}(\epsilon);\epsilon),\alpha^*_{W,1}(\epsilon);\epsilon)$.*

Proof. The proof is in the text. □

Note that for the algorithm to be developed with the level of quality and sold at the price specified in lemma 7, $W_1(q^*_{M,1}(\alpha^*_{W,1}(\epsilon);\epsilon)) \geq 0$ is a sufficient condition but not a necessary one. Indeed, because of H's overestimation bias, it is possible that M has an interest in developing and selling the algorithm even though social welfare would be higher if the algorithm was not sold and no investment was made in its quality. By exploiting H's overestimation bias, M is able to charge H a higher price for the algorithm than the objective value of the information provided by the prediction.

4.1.4 The optimal sharing of liability

In section 4.1.2, we have seen that if H has an overestimation bias ($\epsilon > 0$), when H does not exert any judgment effort, then the choice of a strict responsibility rule of M ($\alpha = 0$) is socially optimal. Moreover, the choice of this liability sharing rule guarantees that H will indeed exert no judgment effort (lemma 1) and that given this strict liability of M ($\alpha = 0$), the quality of the algorithm chosen by M is socially optimal (lemma 2).

These results imply that with a strict liability of M ($\alpha = 0$), the expected social welfare is:

$$\max \left\{0, W_0(q^*_{W,0})\right\} \tag{19}$$

Indeed, if $W_0(q^*_{W,0}) < 0$, then the strict liability of M ($\alpha = 0$) results in M choosing to not develop and distribute the algorithm (*i.e.* M chooses $q = 0$ and $p > p_0(0, 0; \epsilon)$). Conversely, if $W_0(q^*_{W,0}) \geq 0$, then M's strict liability ($\alpha = 0$) ensures that the algorithm will be developed (with a quality level $q^*_{W,0}$) and distributed (at price $p_0(q^*_{W,0}, 0; \epsilon)$). Whether the expected social welfare $W_0(q^*_{W,0})$ is positive or negative, the choices of M coincide with the first best optimum.

In section 4.1.3, we have seen that if H has an overestimation bias ($\epsilon > 0$), when H makes a judgment effort and whatever the liability sharing rule, the quality chosen by M, given by $q^*_{M,1}(\alpha, \epsilon)$, is lower than the socially optimal quality $q^*_{W,1}$, and that this gap increases with respect to the fraction α of the liability that falls on H (lemma 3). Consequently, A will choose to increase M's liability (lemma 4). However, A is constrained by the fact that a higher liability of M (and, in the extreme case, the strict liability of M) may be incompatible with H making a judgment effort (lemma 1). Thus, when the cost of the judgment effort is not too high (*i.e.* $0 \leq k \leq k^*_1(\epsilon)$), liability sharing $\alpha = \alpha^*_{W,1}(\epsilon) \in (0, 1]$ allows A to obtain the highest possible level of quality of the algorithm ($q = q^*_{M,1}(\alpha^*_{W,1}(\epsilon); \epsilon)$) that is compatible with H exerting a judgment effort (lemma 5), although this liability sharing rule is only a second best optimum.

Assuming that $k \leq k^*_1(\epsilon)$, when is the responsibility sharing $\alpha = \alpha^*_{W,1}(\epsilon)$ actually implemented? The policymaker (A) can always choose a strict liability rule of M ($\alpha = 0$), in which case the expected social welfare is given by (19). He will therefore strictly prefer the liability sharing rule $\alpha = \alpha^*_{W,1}(\epsilon)$ if:

$$W_1(q^*_{M,1}(\alpha^*_{W,1}(\epsilon); \epsilon)) > \max \left\{0, W_0(q^*_{W,0})\right\} \tag{20}$$

If A prefers the liability sharing rule $\alpha^*_{W,1}(\epsilon)$ to $\alpha = 0$ (*i.e.* condition (20) is satisfied), then it will also be preferred by M (lemma 6, part (i)): in this case, it is therefore impossible for M to prefer a level of quality of the algorithm higher than $q^*_{M,1}(\alpha^*_{W,1}(\epsilon); \epsilon)$ in order to prevent H from making a judgment effort, or for M to prefer to not develop and distribute the algorithm (lemma 7).

Proposition 1. *(i) If $\epsilon = 0$ (H does not have an overestimation bias), an optimal liability sharing rule is $\alpha = 1$ (strict liability of H). (ii) If $\epsilon > 0$ (H has an overestimation bias), $k \leq k_1^*(\epsilon)$ and (20) is satisfied, then the optimal liability sharing rule is $\alpha = \alpha_{W,1}^*(\epsilon)$. Otherwise, an optimal liability sharing rule is $\alpha = 0$ (strict liability of M).*

Proof. The proof is in the text. □

The intuition of proposition 1 is the following. Let us first consider the case in which H does not have an overestimation bias ($\epsilon = 0$). Her estimation of the quality of the algorithm is correct ($\phi(q) = \tilde{\phi}(q, \epsilon)$). Consequently, the price she is willing to pay for the algorithm perfectly reflects the social value of the information given by the prediction. Thus, even if M does not directly bear the external damage h when a decision $y = R$ is made when $\theta = \theta_2$, he indirectly internalizes the expected damage via the price paid by H, with the consequence that his choice of the quality of the algorithm, as well as his decision to distribute the algorithm, coincide with the first best optimum. This is not true anymore if H suffers from an overestimation bias ($\epsilon > 0$). Indeed, in this case, H's overconfidence in the prediction, which is exploited by M, may discourage H from exerting a judgment and may lead to a lower than optimal level of the quality of the algorithm. Moreover, the price of the prediction is higher than its social value.

The optimal liability sharing rule is generally not unique when $\epsilon = 0$. However, although there may be several optimal liability sharing rules, a strict liability of H is a particularly sensible solution. First, it is a liability rule which is easy to implement for the policymaker. Second, if $\alpha < 1$, then H no longer fully internalizes the external harm. Consequently, for a cost of the judgment effort such that $k > k_H(q, \alpha; 0)$ (with $\alpha < 1$ and for a given level of the quality of the algorithm q), H will not exert a judgment effort after observing a prediction $s_a = \theta_1$, whereas this effort is socially desirable if $k \leq k_H(q, 1; 0)$. Since $\frac{\partial k_H}{\partial \alpha}(q, \alpha; 0) > 0$, for $k \in (k_H(q, \alpha; 0), k_H(q, 1; 0))$, H does not exert a judgment effort whereas it would have been socially desirable for her to do so. If $\alpha = 1$, this situation cannot occur, because the interval of values of k for which H does not make a judgment effort when it would have been socially desirable for her to do so is empty.

4.2 Negligence rule

In this section, we study the efficiency of the negligence rule. The standard cannot be based on the human operator's decision whether to exercise judgment, since judgment is

an inobservable cognitive effort. Therefore, we focus on negligence by the AI manufacturer, where the standard specifies a minimum level of quality of the algorithm, assuming that quality is perfectly observable and verifiable. If M chooses a level of quality higher or equal to the standard, he is not liable in the event of an accident: strict liability of H ($\alpha = 1$) applies. However, if M chooses a level of quality lower than the standard, he his fully liable for the external harm ($\alpha = 0$).

We consider two possible cases, according to the cost of judgment effort. If $k \leq k_H(q^*_{W,1}, 1; \epsilon)$ (respectively $k > k_H(q^*_{W,0}, 1; \epsilon)$), the standard is set to $q^*_{W,1}$ ($q^*_{W,0}$). A quality of the algorithm equal to this standard maximizes the expected social welfare when H exerts (does not exert) a judgment effort.

Proposition 2. *(i) Under the negligence rule with a standard $q^*_{W,0}$, assuming that $k > k_H(q^*_{W,0}, 1; \epsilon)$ and $\pi_0(q^*_{W,0}, 1; \epsilon) \geq 0$, the algorithm will be developed with a level of quality $q^*_{W,0}$ and distributed at price $p_0(q^*_{W,0}, 1; \epsilon)$. H does not exert a judgment effort. (ii) Under the negligence rule with a standard $q^*_{W,1}$, assuming that $k \leq k_H(q^*_{W,1}, 1; \epsilon)$ and $\pi_1(q^*_{W,1}, 1; \epsilon) \geq 0$, the algorithm will be developed with a level of quality $q^*_{W,1}$ and distributed at a price $p_1(q^*_{W,1}, 1; \epsilon)$. H exerts a judgment effort.*

Proof. The proof is in the appendix (section 6.5). □

In the two cases described in proposition 2, a sufficient condition for the algorithm to be actually developed (with a level of quality equal to the standard) and distributed by M is that the expected social welfare must be positive. More specifically, in case (i), M's expected profit can be rewritten:

$$\pi_0(q^*_{W,0}, 1; \epsilon) = W_0(q^*_{W,0}) + \frac{1}{2}\left(\tilde{\phi}(q; \epsilon) - \phi(q)\right) h \geq W_0(q^*_{W,0}) \tag{21}$$

Thus, if $W_0(q^*_{W,0}) \geq 0$, it is better for M to develop and distribute the algorithm than not to do (and obtain an expected profit of 0). In case (ii), following similar reasoning, lemma 6 implies that $\pi_1(q^*_{W,1}, 1; \epsilon) \geq W_1(q^*_{W,1})$. Thus, if $W_1(q^*_{W,1}) \geq 0$, it is better for M to develop and distribute the algorithm.

Note that in both cases (i) and (ii), if H is bias-free ($\epsilon = 0$), M's expected profit is equal to the expected social welfare. Consequently, M will develop and distribute the algorithm only if its expected social value is positive. However, this is no longer true if H suffers from an

overestimation bias ($\epsilon > 0$): M may be willing to develop and distribute the algorithm even though it leads to a loss of social welfare. This is because M takes advantage of H's bias in order to get a rent at her expense.

Note also that $k_H(q^*_{W,0}, 1; \epsilon) < k_H(q^*_{W,1}, 1; \epsilon)$, since $q^*_{W,0} > q^*_{W,1}$ (as shown above with (38)) and $\frac{\partial k_H}{\partial q}(q, 1; \epsilon) < 0$. This implies that for intermediate values of the judgment effort cost (such that $k \in (k_H(q^*_{W,0}, 1; \epsilon), k_H(q^*_{W,1}, 1; \epsilon)]$), A has two options, both of which can be optimal. First, he can choose to set a "high" quality standard $q^*_{W,0}$, with which M will comply. That standard will deter H from making a judgment effort. Second, he can choose to set a "low" quality standard $q^*_{W,1}$, with which M will also comply. That low standard incentivizes H to exert a judgment effort.

Thus, when choosing between the two standards ($q^*_{W,0}$ or $q^*_{W,1}$), the policymaker faces a trade-off. Indeed, the two standards induce different costs (cost of judgment effort and cost of quality), expected gain and expected harm from a risky decision. First, regarding the costs, choosing a standard $q^*_{W,0}$ over a standard $q^*_{W,1}$ saves the cost of the judgment effort, but increases the cost of the investment made by M in the quality of the algorithm ($c(q^*_{W,1}) < c(q^*_{W,0})$). Second, under a $q^*_{W,0}$ standard, H's expected gain from a risky decision (excluding the liability cost) is higher.[29] Third, which standard is associated with the lower expected harm from a risky decision is ambiguous. On the one hand, under a $q^*_{W,0}$ standard, the quality of the algorithm is higher ($q^*_{W,0} > q^*_{W,1}$). Consequently, the probability of an incorrect prediction $s_a = \theta_1 \neq \theta$ (and thus of a decision $y = R$ when $\theta = \theta_2$) is reduced, resulting in a lower expected external harm. On the other hand, under a $q^*_{W,1}$ standard, H makes a judgment effort: a prediction $s_a = \theta_1$ is no longer sufficient to make a decision $y = R$ (it is also necessary to observe a payoff $w = \overline{w}$ in the state of the world $\theta = \theta_1$). Thus, the risky decision is made less often when an incorrect prediction $s_a = \theta_1 \neq \theta$ is observed, reducing the expected external harm.

5 Discussion and concluding remarks

In many fields, we observe an increasingly important role for advisory algorithms to help human decision-making. When these decisions can inflict damage on third parties, it is

[29] This is because, when the gains from a risky decision are $(0, \overline{w})$ and the prediction is $s_a = \theta_1 \neq \theta$, the risky decision is taken nonetheless (since H does not exercise judgment and thus does not observe the gain $w = 0$ associated with θ_1), allowing H to obtain the $\overline{w}$ gain.

necessary to establish a liability rule. In its absence, both the producer of the algorithm and the user may reduce their effort excessively from a social point of view. The question addressed in this paper is then: What types of liability rules should be implemented? We have considered strict liability, sharing of liability and negligence rules. In our framework, a human may buy an algorithm prediction, and, depending on the prediction, may choose to exercise a cognitive effort to assess the gains from making a risky decision. The decision is risky because it may generate damage for a third party who has no direct influence on its occurrence. Indeed, the risk of harm comes from the possibility of a wrong prediction, and may be heightened by the lack of cognitive effort by the human. In order to deal with this issue, we depart on two points from the classical literature on products potentially harmful to third parties (Spence, 1977; Hay and Spier, 2005). First, following Agrawal et al. (2018, 2019a,b), we model the decision-making process by assuming that the algorithm provides a prediction about the state of the world, while the human has the possibility of making a cognitive effort of judgment to better assess the payoffs in each state of the world. Second, we consider that the human may overestimate the reliability of the algorithm's prediction, which is a bias commonly observed in this context.

First, regarding the rules of strict liability and sharing of liability, our main results are the following. Absent any bias, the expected external harm is perfectly internalized by the manufacturer, either through his share of liability, or through the price he charges to the human operator (which is equal to her reservation price for the algorithm, since we assume the manufacturer is a monopolist). Therefore, whatever the sharing of liability (including strict liability of the manufacturer or of the human operator), the manufacturer chooses the level of quality which is socially optimal, and the subjective value (for the operator) of the information provided by the prediction is equivalent to its social value. Strict liability of the human operator provides the additional benefit that it eliminates the risk that she does not exert the judgment effort, while it would have been socially optimal to do so. Thus, in this context, strict liability of the human operator appears to be a sensible rule. However, if the human overestimates the algorithm's accuracy, her reservation price will be higher than the objective value of the prediction. As a result, the manufacturer chooses a level of quality lower than what is socially optimal. Increasing the manufacturer's liability may then be socially beneficial, since it positively impacts the quality chosen by the manufacturer. However, it might also affect the decision of the operator whether to exert a judgment effort. Indeed, a low level of liability for the operator may deter her from making a judgment effort. If that effort is socially desirable, the optimal sharing rule thus involves a trade-off between

these two effects (*i.e.* incentivizing the manufacturer to improve the quality, while ensuring that a judgment effort is made). In the specific case where it is socially optimal that the operator refrains from the judgment, then only a strict liability rule of the manufacturer allows the first best optimum to be reached.

Second, we find that a negligence rule allows the first best optimum to be reached. Under that rule, when the manufacturer complies with the standard, the human operator is liable for the entirety of the external harm. Assuming that the standard is set to the socially optimal level of quality, we have shown that the manufacturer complies with it (and thus chooses the socially optimal quality). Moreover, the operator uses the algorithm in a socially optimal fashion, because she perfectly internalizes the expected external harm. We also show that the policymaker, when he chooses the standard, faces a trade-off between improving the quality of the prediction and making the operator exert her judgment effort.

Although our results show that an appropriately designed negligence rule averts some of the limitations of the strict liability and the sharing of liability rules, it may be difficult to implement in practice, especially when applied to external damages resulting from an incorrect AI prediction.[30] We see mainly three limitations that may prevent a negligence rule from being socially optimal.

A first limitation concerning the use of a negligence rule is that the socially optimal level of the standard may be difficult to determine for the public authority, since it requires knowledge of both the effect of a greater investment in quality on the reliability of the algorithm's prediction, and the costs and benefits of increasing that quality. As a result, other liability rules may do better, as "manufacturers are likely to be better informed about the feasibility of product modifications than regulators" (Hay and Spier, 1997). The consequences of a poorly designed negligence rule can be substantial. Assume for instance that the standard is set too low. The quality of the algorithm chosen by the AI manufacturer will be inefficiently low, and the human's overestimation bias will prevent her from fully perceiving that low quality. Conversely, if the standard is set too high, the AI manufacturer may prefer to not comply with the standard by choosing a lower level of quality. In this case, the liability rests fully with the AI manufacturer and the operator, assuming she (correctly) expects non-compliance with the standard by the manufacturer, will never exercise her judgment effort, even in cases where this would be socially desirable.

[30] For instance, Shavell (2009) states that mistakes related to the level of the standard is "a problem that may be of general significance for [...] firms using new technology".

A second limitation is that the AI manufacturer may misperceive the level of the standard that is enforced by courts, either due to the vagueness of the terms that are used to formulate the standard, or due to the uncertainty toward the way the court will interpret that formulation. It is especially true if the standard can only be given in very broad terms, due to the practical impossibility of specifying exactly a level of quality 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 was actually made by the AI manufacturer to improve the algorithm, and hence to determine accurately whether or not the manufacturer was negligent. By way of illustration, consider the example of an incorrect prediction provided by a deep learning algorithm. These algorithms are often considered as "black boxes" since it may be very difficult, even for their creators, to identify the weight attached to each determinant (the input given to the algorithm) and how these determinants relate to each other in order to shape the algorithm's prediction. Unless the code of the algorithm is closely examined (and sometimes even in this case), it may be difficult to determine whether the poor quality of a prediction results from negligence in the algorithm development process (in which case the AI manufacturer is at fault), from biased training data (in which case the negligence may be assigned to other actors and potentially to the human operator), or from some other misuse of the algorithm. Even if the code of the algorithm is closely examined, the risk of a mistake (a finding of negligence of the AI manufacturer when it is not the case, or conversely) still exists. As is well known from the law and economics literature (*e.g.* Shavell, 1987), mistakes in the finding of 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 quality.

Although, due to these limitations, a negligence rule can be tricky to implement efficiently in practice, the applicability of a liability sharing rule is not straightforward either. Indeed, designing a very specific apportionment of liability (like the one that achieves the highest possible level of quality while ensuring that the user exercises a judgment effort) may be rather challenging to set up. Moreover, the optimal sharing will generally depend on the particular context to which it is applied. However, because the optimal sharing rule relies only on the operator's willingness to exert her judgment effort, we may expect that determining that optimal sharing rule is less information-intensive than determining an optimal standard of quality under a negligence rule. Finally, even though there remain some uncer-

tainties regarding the exact sharing of liability that should apply, the trade-off highlighted in section 4.1 between, on the one hand, incentivizing the AI manufacturer to improve the quality (by increasing his share of liability) and, on the other hand, incentivizing the human operator to exert a judgment effort (assuming this effort is socially beneficial) still applies.

To conclude, in the context we study, both rules (negligence and sharing of liability) have specific strengths and weaknesses, and the comparison of these rules is still up for debate. A few extensions of our model may be considered. For instance, 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 consumer (Bienenstock, 2016). In the latter case, in the context of decision-making with an advisory algorithm, the consumer tends to overestimate the quality of the product. Therefore, it is not rational for a monopoly to educate the consumer. However, this could be the case in an oligopoly context: a manufacturer might be willing to correct the consumers' perception of the quality of its competitors' products. Such an extension would be relevant in our setting, in which it would be worth investigating how liability rules might prompt AI manufacturers to invest in consumer education. These extensions are left for future research.

6 Appendix

6.1 Proof of lemma 3

Comparing the FOCs for $q^*_{M,1}(\alpha;\epsilon)$ and $q^*_{W,1}$, we have:

$$\frac{\partial \pi_1}{\partial q}(q,\alpha;\epsilon) - \frac{\partial W_1}{\partial q}(q) = \frac{v}{2}\left(\frac{\partial \tilde{\phi}}{\partial q}(q;\epsilon) - \frac{\partial \phi}{\partial q}(q)\right)((1-v)\overline{w} + \alpha h)$$
$$< (=)\ 0\ \forall \alpha \in [0,1] \text{ if } \epsilon > (=)\ 0 \tag{22}$$

Moreover, we have:

$$\frac{\partial^2 \pi_1}{\partial q \partial \alpha}(q,\alpha;\epsilon) = \frac{v}{2}\left(\frac{\partial \tilde{\phi}}{\partial q}(q;\epsilon) - \frac{\partial \phi}{\partial q}(q)\right) h < (=)\ 0 \text{ if } \epsilon > (=)\ 0 \tag{23}$$

Which implies, from the implicit function theorem, that $\frac{\partial q^*_{M,1}}{\partial \alpha}(\alpha;\epsilon) < 0$.

6.2 Proof of lemma 4

We have:

$$\frac{\mathrm{d}W_1(q^*_{M,1}(\alpha;\epsilon))}{\mathrm{d}\alpha} = \frac{\partial q^*_{M,1}}{\partial \alpha}(\alpha;\epsilon)\left(\frac{v}{2}\frac{\partial \phi}{\partial q}(q^*_{M,1}(\alpha;\epsilon))((1-v)\overline{w}+h) - c'(q^*_{M,1}(\alpha;\epsilon))\right) \tag{24}$$

We know from lemma 3 that $\frac{\partial q^*_{M,1}}{\partial \alpha}(\alpha;\epsilon) < 0$ if $\epsilon > 0$. Thus, $W_1(q^*_{M,1}(\alpha;\epsilon))$ is strictly decreasing with α (*i.e.* $\frac{\mathrm{d}W_1(q^*_{M,1}(\alpha;\epsilon))}{\mathrm{d}\alpha} < 0$) if and only if:

$$c'(q^*_{M,1}(\alpha;\epsilon)) > \frac{v}{2}\frac{\partial \phi}{\partial q}(q^*_{M,1}(\alpha;\epsilon))((1-v)\overline{w}+h) \tag{25}$$

From the FOC of $q^*_{M,1}(\alpha;\epsilon)$, we have:

$$c'(q^*_{M,1}(\alpha;\epsilon)) = \frac{v}{2}\left(\frac{\partial \tilde{\phi}}{\partial q}(q^*_{M,1}(\alpha;\epsilon);\epsilon)((1-v)\overline{w}+\alpha h) + \frac{\partial \phi}{\partial q}(q^*_{M,1}(\alpha;\epsilon))(1-\alpha)h\right) \tag{26}$$

By substituting (26) in (25), we can rewrite (25) as:

$$\frac{v}{2}\left(\frac{\partial \tilde{\phi}}{\partial q}(q^*_{M,1}(\alpha;\epsilon);\epsilon)((1-v)\overline{w}+\alpha h) + \frac{\partial \phi}{\partial q}(q^*_{M,1}(\alpha;\epsilon))(1-\alpha)h\right) > \frac{v}{2}\frac{\partial \phi}{\partial q}(q^*_{M,1}(\alpha;\epsilon))((1-v)\overline{w}+h) \tag{27}$$

If $\epsilon > 0$, this condition is always satisfied. Following similar reasoning, if $\epsilon = 0$, then $\frac{\mathrm{d}W_1(q^*_{M,1}(\alpha;\epsilon))}{\mathrm{d}\alpha} = 0$.

6.3 Proof of lemma 5

In a first part of this proof, we prove that if $k \to 0$ (resp. $k \to +\infty$), then $\alpha^*_{W,1}(\epsilon) \in (0,1]$ (resp. $\alpha^*_{W,1}(\epsilon) > 1$).

From (3) and (18), we have:

$$\alpha^*_{W,1}(\epsilon) = \frac{v\overline{w} + \frac{k}{(1-v)(1-\tilde{\phi}(q^*_{M,1}(\alpha^*_{W,1}(\epsilon);\epsilon);\epsilon))}}{h} \tag{28}$$

Moreover, $q^*_{M,1}(\alpha;\epsilon)$ does not depend on k since $\frac{\partial^2 \pi_1}{\partial q \partial k}(q,\alpha;\epsilon) = 0$. Thus:

$$\lim_{k\to 0}\left(\frac{v\overline{w} + \frac{k}{(1-v)(1-\tilde{\phi}(q^*_{M,1}(\alpha^*_{W,1}(\epsilon);\epsilon);\epsilon))}}{h}\right) = \frac{v\overline{w}}{h} \in (0,1) \tag{29}$$

And:

$$\lim_{k\to \infty}\left(\frac{v\overline{w} + \frac{k}{(1-v)(1-\tilde{\phi}(q^*_{M,1}(\alpha^*_{W,1}(\epsilon);\epsilon);\epsilon))}}{h}\right) = +\infty \tag{30}$$

In a second part of this proof, we prove that $\alpha^*_{W,1}(\epsilon)$ is monotonically increasing with respect to k by using the implicit function theorem.

Let us define:

$$L(\alpha,k) = k_H(q^*_{M,1}(\alpha;\epsilon),\alpha;\epsilon) - k \tag{31}$$

Thus, (18) is equivalent to $L(\alpha,k) = 0$. We have:

$$\frac{\partial L}{\partial k}(\alpha,k) = -1 \tag{32}$$

$$\frac{\partial L}{\partial \alpha}(\alpha,k) = (1-v)\left((1-\tilde{\phi}(q^*_{M,1}(\alpha;\epsilon);\epsilon))h - (\alpha h - v\overline{w})\frac{\partial \tilde{\phi}}{\partial q}(q^*_{M,1}(\alpha;\epsilon);\epsilon)\frac{\partial q^*_{M,1}}{\partial \alpha}(\alpha;\epsilon)\right) \tag{33}$$

Thus, from the implicit function theorem, the sign of the effect of k on $\alpha^*_{W,1}(\epsilon)$ is the same as the sign of:

$$(1-\tilde{\phi}(q^*_{M,1}(\alpha^*_{W,1}(\epsilon);\epsilon);\epsilon))h - (\alpha^*_{W,1}(\epsilon)h - v\overline{w})\frac{\partial \tilde{\phi}}{\partial q}(q^*_{M,1}(\alpha^*_{W,1}(\epsilon);\epsilon);\epsilon)\frac{\partial q^*_{M,1}}{\partial \alpha}(\alpha^*_{W,1}(\epsilon);\epsilon) \tag{34}$$

By substituting (28) in (34), we can rewrite (34) as:

$$(1-\tilde{\phi}(q^*_{M,1}(\alpha^*_{W,1}(\epsilon);\epsilon);\epsilon))h - \frac{k}{(1-v)(1-\tilde{\phi}(q^*_{M,1}(\alpha^*_{W,1}(\epsilon);\epsilon);\epsilon))}\frac{\partial \tilde{\phi}}{\partial q}(q^*_{M,1}(\alpha^*_{W,1}(\epsilon);\epsilon);\epsilon)\frac{\partial q^*_{M,1}}{\partial \alpha}(\alpha^*_{W,1}(\epsilon);\epsilon) \tag{35}$$

Since $\frac{\partial q^*_{M,1}}{\partial \alpha}(\alpha^*_{W,1}(\epsilon);\epsilon) \leq 0$, (35) is positive : $\alpha^*_{W,1}(\epsilon)$ is monotonically increasing with respect to k.

From the two parts of this proof, we can deduce that there exists a unique threshold value

of k, which we denote $k_1^*(\epsilon)$, such that $\alpha_{W,1}^*(\epsilon) \in (0,1]$ (resp. $\alpha_{W,1}^*(\epsilon) > 1$) if $k \leq k_1^*(\epsilon)$ (resp. if $k > k_1^*(\epsilon)$).

6.4 Proof of lemma 6

We have:

$$\frac{\partial \pi_1}{\partial \alpha}(q, \alpha; \epsilon) = \frac{v}{2}\left(\tilde{\phi}(q;\epsilon) - \phi(q)\right) h > (=)\ 0 \text{ if } \epsilon > (=)\ 0 \tag{36}$$

And:

$$\pi_1(q, 0; \epsilon) - W_1(q) = \frac{v}{2}(1-v)(\tilde{\phi}(q;\epsilon) - \phi(q))\overline{w} > (=)\ 0 \text{ if } \epsilon > (=)\ 0 \tag{37}$$

Thus, if $\epsilon > 0$, whatever the quality of the algorithm (q) and the liability sharing rule (α), M's expected profit is higher than the expected social welfare.

6.5 Proof of proposition 2

First, let us consider the case where $k \leq k_H(q_{W,1}^*, 1; \epsilon)$. Assume that the development of the algorithm with a quality level $q_{W,1}^*$, and its distribution at a price $p_1(q_{W,1}^*, 1; \epsilon)$, allows M to obtain a positive expected profit ($\pi_1(q_{W,1}^*, 1; \epsilon) \geq 0$). We show by contradiction that M complies with the standard $q_{W,1}^*$. Assume that M chooses not to comply with the standard. In other words, M chooses a quality level $q < q_{W,1}^*$. Because the standard is not met, M is strictly liable ($\alpha = 0$). As a consequence, H does not exert a judgment effort (lemma 1). Anticipating that, the quality level chosen by M will be $q_{M,0}^*(0;\epsilon) = q_{W,0}^*$ (lemma 2). However, we have:

$$\frac{\partial W_0}{\partial q}(q) - \frac{\partial W_1}{\partial q}(q) = \frac{1}{2}(1-v)(h - v\overline{w})\phi'(q) > 0\ \forall q \tag{38}$$

This implies that $q_{W,0}^* > q_{W,1}^*$, and thus that M will choose a level of quality $q_{M,0}^*(0;\epsilon) > q_{W,1}^*$, which contradicts the assumption that M chooses $q < q_{W,1}^*$. We conclude that M will comply with the standard.

Will M choose to adopt a level of quality higher than the standard $q_{W,1}^*$? If M chooses a level of quality higher than the standard ($q \geq q_{W,1}^*$), H is fully liable for the external harm ($\alpha = 1$). In these circumstances, M's expected profit is decreasing with the level of quality

of the algorithm:[31]

$$\frac{\partial \pi_1}{\partial q}(q, 1; \epsilon) < 0 \text{ if } q > q^*_{W,1} \tag{39}$$

Therefore, consistently with the standard reasoning with respect to negligence rules, M chooses the level of quality which just meets the standard. Thus, we conclude that when $k \leq k_H(q^*_{W,1}, 1; \epsilon)$ and if a standard $q^*_{W,1}$ is (perfectly) enforced, M chooses a level of quality equal to this standard.

Second, let us consider the case where $k > k_H(q^*_{W,0}, 1; \epsilon)$. Assume that $\pi_0(q^*_{W,0}, 1; \epsilon) \geq 0$. Using similar reasoning to that above, we show by contradiction that M complies with the standard $q^*_{W,0}$. Assume that M does not comply, by choosing $q < q^*_{W,0}$. As a consequence, M is strictly liable ($\alpha = 0$) and, from lemma 1, H does not exert a judgment effort. Anticipating that, the level of quality will be $q^*_{M,0}(0; \epsilon) = q^*_{W,0}$ (lemma 2), which contradicts the assumption that M chooses $q < q^*_{W,0}$. Also, it can easily be shown that M will not choose a level of quality higher than $q^*_{W,0}$. Thus, we conclude that when $k > k_H(q^*_{W,0}, 1; \epsilon)$ and if a standard $q^*_{W,0}$ is enforced, M chooses once again a level of quality which is equal to the standard.

References

Abrardi, L., Cambini, C., and Rondi, L. (2021). Artificial intelligence, firms and consumer behavior: A survey. *Journal of Economic Surveys*.

Aghion, P. and Tirole, J. (1997). Formal and real authority in organizations. *Journal of Political Economy*, 105(1):1–29.

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.

[31] This result comes from lemma 3 and the fact that $\frac{\partial^2 \pi_1}{\partial q^2}(q, 1; \epsilon) < 0$.

Athey, S. C., Bryan, K. A., and Gans, J. S. (2020). The allocation of decision authority to human and artificial intelligence. In *AEA Papers and Proceedings*, volume 110, pages 80–84.

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.

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*.

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

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., and Manna, E. (2021). Car accidents in the age of robots. *International Review of Law and Economics. forthcoming*.

Ebers, M. (2021). Liability for artificial intelligence and eu consumer law. *J. Intell. Prop. Info. Tech. & Elec. Com. L.*, 12:204.

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

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. (2021a). Liability for robots i: Legal challenges. *Available at SSRN 3939477*.

Guerra, A., Parisi, F., and Pi, D. (2021b). Liability for robots ii: An economic analysis. *Available at SSRN 3939486*.

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.

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.

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.

Liang, A., Lu, J., and Mu, X. (2021). Algorithmic design: Fairness versus accuracy. *arXiv preprint arXiv:2112.09975*.

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.

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.

Rambachan, A., Kleinberg, J., Mullainathan, S., and Ludwig, J. (2020). An economic approach to regulating algorithms. Technical report, National Bureau of Economic Research.

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

Shavell, S. (2009). *Foundations of economic analysis of 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.

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.

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.

Zhang, Y., Liao, Q. V., and Bellamy, R. K. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305.