Artificial intelligence, inattention and liability rules

Marie Obidzinski, Yves Oytana

CENTRE DE RECHERCHE
SUR LES STRATÉGIES ÉCONOMIQUES

Artificial intelligence, inattention and liability rules

Marie Obidzinski, Yves Oytana

February 2024

Working paper No. 2024 – 08

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

Artificial intelligence, inattention and liability rules

Marie Obidzinski,* Yves Oytana†

February 9, 2024

Abstract

We characterize the socially optimal liability sharing rule in a situation where a manufacturer develops an artificial intelligence (AI) system that is then used by a human operator (or user). First, the manufacturer invests to increase the autonomy of the AI (*i.e.*, the set of situations that the AI can handle without human intervention) and sets a selling price. The user then decides whether or not to buy the AI. Since the autonomy of the AI remains limited, the human operator must sometimes intervene even when the AI is in use. Our main assumption is that users are subject to behavioral inattention. Behavioral inattention reduces the effectiveness of user intervention and increases the expected harm. Only some users are aware of their own attentional limits. Under the assumption that AI outperforms users, we show that policymakers may face a trade-off when choosing how to allocate liability between the manufacturer and the user. Indeed, the manufacturer may underinvest in the autonomy of the AI. If this is the case, the policymaker can incentivize the latter to invest more by increasing his share of liability. On the other hand, increasing the liability of the manufacturer may come at the cost of slowing down the diffusion of AI technology.

Keywords: liability rules, artificial intelligence, inattention.

JEL classification: K4.

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

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

1 Introduction

Motivation. Autonomous artificial intelligence is a type of artificial intelligence (AI) that uses the latest technological advances to enable devices to perform tasks independently of the human operator. So-called "autonomous" vehicles are an example of the development of this type of AI. However, at the current stage of research and development, these vehicles are not yet fully autonomous. This means that the user must be able to stay alert enough to respond quickly to system alerts and to regain control in situations that the AI cannot handle. This requires the user to maintain a high level of attention, sometimes for long periods of time.

This problem of maintaining attention has long been recognized in computer science (at the intersection with cognitive science), along with overreliance on automation. Cognitive limitations make it difficult for a human to maintain visual attention (Bainbridge, 1983), while overreliance on automation (or "automation bias") is the tendency to place more weight or trust in machine output than in our own human capabilities.[1] Both of these dimensions are worth considering when deciding how to allocate liability in the event of an accident involving an automated system. In this paper, we have chosen to focus on the attention issue that arises with increased automation.[2]

Lack of attention when using an "autonomous" device can create a non-negligible risk of accident. The first fatal accident involving a car with an autopilot system occurred in 2016, due to a defect in the car's sensors. The driver was unable to regain control of the car quickly enough.[3] The automaker said that while "autopilot is getting better all the time, it is not perfect and still requires the driver to stay alert."[4] In 2019, for the first time, a driver was sued after

[1] Zerilli et al. (2019) generally refer to the difficulties of the operator in a human-machine control loop as the "control problem." The authors identify several types of difficulties that can arise between a human operator and a machine, such as the "attention problem" and the "attitude problem." The latter refers to the automation bias, while the former refers to the vigilance problem (see also Parasuraman and Riley, 1997; Cummings, 2017; Alberdi et al., 2009).

[2] In a companion paper, we focus on the automation bias (Obidzinski and Oytana, 2022).

[3] This problem has also been identified in other areas, such as aviation. Projects based on artificial intelligence are being developed to increase the autonomy of aircraft, with the common goal of at least partially replacing co-pilots. For example, with the completion of the Dragonfly project by a subsidiary of Airbus, aircraft could take off and land automatically. However, in the event of a problem, these systems require the pilot to regain control.

[4] "Autopilot is getting better all the time, but it is not perfect and still requires the driver to stay alert." Source: The Guardian, July 1, 2016, https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk. Some fatal accidents have also involved third parties, such as pedestrians. Source: https://www.npr.org/2022/01/18/1073857310/tesla-autopilot-crash-charges.

causing an accident by misusing an automated driving system. The driver crashed his car into another vehicle, killing two passengers, and was prosecuted by California authorities.[5] Similar issues can arise with decision support algorithms. For example, in radiology, algorithms provide probabilities that a patient is positive for a pathology. To make a diagnosis, radiologists may use this information as well as their own expertise and additional contextual information on which the algorithm was not trained (*e.g.*, for privacy reasons). Therefore, the operator should remain vigilant, even when the algorithm makes very accurate predictions.

In terms of efficiency, these difficulties do not mean that it would be socially beneficial to eliminate these algorithms. On the contrary, the use of such systems generally reduces the overall risk of accidents, although the type of risk that materializes in practice is often of a different nature.[6] It is important to note, however, that the allocation of liability between the AI manufacturer and the user has important welfare implications, by affecting their incentives. Indeed, liability will affect the manufacturer's investment in developing the AI and improving its autonomy, as well as the user's level of attention when using the AI and her decision to purchase the AI in the first place.

Research question and main assumptions. In this context, we investigate on the optimal allocation of liability between the AI manufacturer and the AI user. We develop a framework in which an AI manufacturer chooses, through a costly investment, both the degree of autonomy of the AI system and a selling price. The human user decides whether or not to buy the automated system. If he decides not to buy the system, his intervention is always necessary. Conversely, if he decides to buy the AI, his intervention is only necessary if a situation arises that the AI cannot handle (the probability of such a situation occurring decreases with the degree of autonomy of the AI). In the case of human intervention, the user chooses an action that may cause some harm if it is not appropriate. The expected harm increases when the user is subject to a higher degree of behavioral inattention, which is modeled following Gabaix (2019). This framework has

[5] "The National Highway Traffic Safety Administration (NHTSA) and the National Transportation Safety Board (NTSB) have been investigating the widespread misuse of autopilot by drivers, whose overconfidence and inattention have been blamed for several crashes, including fatal ones. In one crash report, the NTSB referred to the misuse as automation complacency." Source : National public radio, January 18, 2022, "A Tesla driver is charged in a crash involving Autopilot that killed 2 people", https://www.npr.org/2022/01/18/1073857310/tesla-autopilot-crash-charges.

[6] An autopilot may be the cause of an accident that would have been easily avoided by a human driver, and vice versa.

the advantage of providing a tractable way to model different degrees of behavioral inattention. Another behavioral assumption we make is that only a subset of users (the "sophisticated" users) are fully aware that their inattention can cause harm. As Armstrong and Vickers (2012) pointed out in another context,[7] some users may not account for the costs of their inattention. In our model, inattention can be interpreted as the user being overly optimistic either about the degree of autonomy of the AI, or about their ability to quickly mobilize their attention to make a damage-avoidance decision.[8]

Main results. In our baseline model, we show that the AI manufacturer's investment in improving the AI's autonomy is insufficient if users are subject to behavioral inattention. The manufacturer can be incentivized to invest more by increasing his share of liability. However, increasing his liability share can have significant social costs. First, it may slow down the diffusion of AI (AI will only be purchased by sophisticated users). Second, in extreme cases, the AI manufacturer may even stop developing and marketing the AI system (which is socially harmful, since its use is assumed to reduce the expected harm). In an extension of the model, we assume that the user can manage higher levels of attention through a costly cognitive effort. We show that the cognitive effort chosen by sophisticated users increases with their share of liability, adding a new element to the trade-off identified in our baseline model.

Related literature. Our article is related to the literature on liability for defective products (Landes and Posner, 1985; Daughety and Reinganum, 2013; Hay and Spier, 2005). Specifically, Hay and Spier (2005) propose a bilateral care model in which, if consumer wealth is high enough, consumer-only liability is socially optimal. Indeed, this rule leads consumers to fully internalize the expected harm and to take socially optimal precautions. Moreover, competition ensures that the level of safety and the quantity produced are socially optimal. Inefficiencies arise, however, when consumer wealth is so low that she is unable to pay for the entire harm. In this case, a "residual manufacturer liability" (the manufacturer has to pay the shortfall of damages

[7] Armstrong and Vickers (2012) assume that some individuals are overly optimistic about their ability to avoid an overdraft. These individuals do not properly take into account the fees charged by their bank when an overdraft occurs.

[8] In O'Donoghue and Rabin (1999), "naive" individuals do not anticipate their self-control problems in the future, unlike "sophisticated" individuals. This lack of self-control and the bias of naive individuals help to explain procrastination behavior. In our case, individuals are naive in the sense that they fail to anticipate their difficulty in maintaining a sustained level of attention.

not paid by the consumer) is a second-best solution and limits the inefficiencies arising from the judgment proof problem. The contribution of Hay and Spier (2005) provides a rationale for implementing a sharing of liability between the user of a product and the manufacturer of that product.

Other papers find that liability sharing between the producer and the consumer is socially optimal, but with a rationale related to consumer cognitive biases (Friehe et al., 2020; Obidzinski and Oytana, 2022), as it is the case in the present paper. However, the specific behavioral patterns they consider (present bias and automation bias) are of a different nature than the one considered here. Rather, we focus on attentional issues (*i.e.*, behavioral inattention *à la* Gabaix, 2019) that arise when using semi-autonomous or advisory algorithms. This framework allows for effects of a different nature to emerge when compared to Friehe et al. (2020) and Obidzinski and Oytana (2022).

Several authors have focused on liability rules in the specific case of autonomous vehicles (Shavell, 2020; Talley, 2019; De Chiara et al., 2021; Dawid and Muehlheusser, 2022). Although we also discuss liability rules in a similar context, these contributions are situated within the classical framework of economic analysis of civil liability, which does not take into account possible behavioral biases of human users.

The remainder of the paper is organized as follows. In Section 2, we present a simple example to illustrate the trade-off the policymaker faces when choosing the liability rule. Section 3 presents the baseline model and characterizes the optimal liability sharing rule. In Section 4, we extend the model to the case where the human user can increase her attention by exerting a costly cognitive effort. Section 5 concludes.

2 Manufacturer's strict liability, AI's autonomy and diffusion: An example

In this section, we introduce a simple example to illustrate the main effects at play. Suppose there is an AI system (*e.g.*, an autonomous car) with two possible levels of autonomy: high and low. When the AI's autonomy is high (low, respectively), the user must intervene with

a probability of 0.2 (0.6, respectively). Moreover, if the user does not use the AI provided by the manufacturer (*e.g.*, the human uses a car without a self-driving system), the intervention probability is 1. An intervention costs 40 to the human user. We assume that as the intervention probability decreases, the level of attention also decreases, resulting in a higher level of harm in case of an intervention, as depicted in Table 1. Furthermore, the cost of developing the AI system increases with its level of autonomy.

AI autonomy	Intervention probability	Cost of intervention	Expected harm (if intervention)	Development cost
None	1	40	10	0
Low	0.6	40	20	5
High	0.2	40	30	25

Table 1

The expected social cost is defined as the sum of the development cost and the expected cost of human intervention (including the expected harm). For example, in the case of an AI with a low level of autonomy, the expected social cost is: $5 + 0.6 \times (20 + 40) = 41$. Table 2 shows the expected social costs for each level of autonomy. We observe that a high degree of autonomy of the AI minimizes the expected social costs and is therefore desirable.

AI autonomy	Intervention probability	Expected social cost
None	1	50
Low	0.6	41
High	0.2	39

Table 2

We consider two types of human users: naive and sophisticated. Unlike the sophisticated users, naive users do not consider the consequences of inattention, and thus the expected harm of interventions. As a result, the willingness to pay for the algorithm will be lower for the naive users. The proportion of each type of user is given. In our example, assume that 25% of users are naive, while the remaining 75% are sophisticated.

The manufacturer can sell the AI at a "low" price, so that all types of users buy the AI,[9]

[9]The "low" price is equal to the willingness to pay of the naive user (who underestimates the benefits of the AI). More precisely, the low price is the expected cost saved by not intervening, *i.e.*, the probability of not intervening (0.8 for a high autonomy AI and 0.4 for a low autonomy AI) times the cost of intervening (40).

or at a "high" price, so that only sophisticated users buy it.[10] Consider the case where the manufacturer sets a low price (note that the expected social cost is lower in this case). The price depends on the level of autonomy of the AI. Specifically, the price equals $0.4 \times 40 = 16$ for a low autonomy AI, and $0.8 \times 40 = 32$ for a high autonomy AI, where 0.4 and 0.8 are the probabilities for the user *not* to intervene under respectively a high and a low level of autonomy.

What is the level of autonomy chosen by the manufacturer when the AI is sold at the low price? First, assume that the user bears all the harm (no manufacturer liability). Table 3 shows that the manufacturer gets the highest expected profit by choosing a low level of AI autonomy.[11] If we switch to a strict liability rule (the manufacturer must pay for all the harm), table 3 now shows that the manufacturer receives the highest expected profit by choosing a high level of AI autonomy.[12] Thus, a strict liability rule incentivizes the manufacturer to choose the socially optimal level of autonomy (that is a high level of autonomy), *given that the AI is sold at the low price.*

AI autonomy	Expected profit (no liability)	Expected profit (strict liability)
None	0	0
Low	11	−1
High	7	1

Table 3

However, under strict liability, the manufacturer might be tempted to deviate by choosing a higher price. Indeed, for a high level of AI autonomy, the manufacturer's profit from selling the AI at a high price increases from 1 to 2:[13] strict manufacturer liability encourages manufacturers to choose a high price, thereby excluding some consumers. Therefore, from a policy perspective, there may be a trade-off between incentivizing the manufacturer to increase his effort to develop the autonomy of the AI on the one hand, and promoting the diffusion of the AI technology (by

[10] The "high" price is equal to the willingness to pay of the sophisticated user. It is defined as the user's savings relative to the expected cost of the intervention and the expected harm.

[11] Under no liability, the expected profit is equal to the price minus the development cost. Thus, if AI autonomy is low (high, respectively), it is given by $16 - 5 = 11$ ($32 - 25 = 7$, respectively).

[12] Under strict liability, the expected profit includes the manufacturer's expected liability cost. Thus, if AI autonomy is low (high, respectively), it is given by $16 - 0.6 \times 20 - 5 = -1$ ($32 - 0.2 \times 30 - 25 = 1$, respectively).

[13] Sophisticated users are willing to pay a price $0.8 \times 40 + 10 = 42$. The manufacturer's expected profit is then $0.75 \times (42 - 0.2 \times 30) - 25 = 2$.

not excluding some consumers from its use) on the other hand. Also related to the issue of AI diffusion, table 3 shows that the expected profit of the manufacturer is lower under strict liability than under no liability. This implies that in some cases (*e.g.*, if the development costs are slightly higher than in our example), switching from no liability to strict liability may induce the manufacturer to forgo developing the AI.[14]

The following section formalizes the previous findings.

3 The Model

In this section, we develop a product liability framework where the product under consideration is an AI algorithm (hereafter AI) and the consumer is the human user of the AI.

3.1 The cost of user intervention

In a proportion $\pi \in [0,1)$ of situations, the AI works effectively without human intervention. However, in a proportion $1-\pi$ of situations, the AI does not act autonomously: human intervention is then required. The parameter π is thus interpreted as an index of the degree of autonomy of the AI. When the AI is not used, human intervention is always required. The cost of developing an AI with a degree of autonomy of π is $c(\pi)$, where $c'(0) = 0$, $c'(\pi) \geq 0$ and $c''(\pi) > 0$. There are some fixed costs ($c(0) > 0$). These hypotheses on $c(\pi)$ capture the idea that increasing the autonomy of the AI is costly.

If the user has to intervene, she incurs a cost $k > 0$.[15] In fact, the need for human intervention may imply a cognitive and/or physical effort on the part of the user to deal with the situation not handled by the AI. Thus, if she decides not to buy the AI, she always bears the cost k. Conversely, if she buys the AI, she bears the cost k with probability $(1-\pi)$ (which is the probability that the AI will be confronted with a case it cannot handle autonomously).

The main assumptions of our model are that human users may be inattentive in the sense of Gabaix (2019), and that only some users are aware of this inattention problem. We use the same

[14] More specifically, the development costs should be such that the expected profit of the manufacturer is always negative if he develops an AI under strict liability, while it may be positive under no liability.

[15] This cost is independent of the parameter m and the type of the user (naive or sophisticated).

framework as Gabaix (2019) to model behavioral inattention. In the case of an intervention, the user has to choose an action a. She knows that the appropriate action is the realization of a random variable X, distributed over the support $[\underline{x}, \overline{x}]$ according to the density function $f(x)$. However, because she is subject to behavioral inattention, her action tends to be biased toward the mean x^d of this distribution, which is used as a "default value" when she is not fully attentive. Specifically, we assume that the weight given to the appropriate action m, which can be interpreted as the user's level of attention, is a decreasing function of the AI's level of autonomy π.

To illustrate, consider self-driving cars with conditional automation.[16] The driver must sometimes intervene to respond appropriately to a request issued by the car's AI. The probability of a request, and therefore of intervention, decreases as the car's level of autonomy increases. Suppose the autonomy increases after implementing a better AI engine (higher π). We can expect that the user will find it more difficult (and less useful, see section 4) to maintain his level of attention. As a result, the user's level of attention m will decrease. We formalize this idea with the following assumption:

Assumption 1. *The user's attention level is such that (i)* $m'(\pi) < 0$ *and (ii)* $m(0) > 0$.

Part (i) of Assumption 1 states that the level of attention decreases with the degree of autonomy of the AI. Part (ii) of Assumption 1 means that when the user is not using an AI, his level of attention is strictly positive.[17]

Since $m(\pi)$ is the weight the user assigns to x relative to x^d, the user is subject to behavioral inattention if $m(\pi) < 1$. The appropriate action, as perceived by the human user for a given level of attention m, is :

$$x^s(m) = mx + (1 - m)x^d \tag{1}$$

If the user has to intervene, she chooses an action a and the harm $h(a, x)$ is realized. This harm increases with the distance between the user's chosen action (a) and the appropriate action (x).[18]

[16] According to the Society of Automotive Engineers classification, conditional automation is Level 3 automation.

[17] In an extension of the model (Section 4), we will endogenize the user's choice of attention.

[18] For example, the damage function can be quadratic, with $h(a, x) = \frac{1}{2}(a - x)^2$. Other specifications can be

The user is liable for a part α of the harm, and the manufacturer for the remaining part $(1-\alpha)$. The user chooses a to minimize her expected liability cost (which is equivalent to minimizing the harm $h(a,x)$):

$$a^s(x;m(\pi)) = \arg\min_a \alpha h(a, x^s(m(\pi))) = m(\pi)x + (1-m(\pi))x^d \tag{2}$$

Given this decision rule,[19] the expected damage when using the AI and for a given level of attention m, is given by:

$$(1-\pi)H(m(\pi)) \tag{3}$$

with:

$$H(m) = \int_{\underline{x}}^{\overline{x}} f(x)h(a^s(x;m),x)dx \tag{4}$$

We make the following assumption:

Assumption 2. *The harm function $h(a,x)$ is such that (i) $H(1)=0$ and (ii) $H'(m)<0$.*

Part (i) of Assumption 2 states that if the user pays maximum attention, she will always choose the appropriate action. As a result, no harm will be done. Part (ii) of Assumption 2 means that as the level of attention increases, the expected harm decreases.[20] Unless otherwise stated, we assume in the following that users suffer from some degree of behavioral inattention, with $m(\pi)<1 \forall \pi$, which implies that $H(m(\pi))>0 \forall \pi$.

There are three players in our model: a policymaker, an AI manufacturer (he), and a human user (she). In a first stage, the policymaker chooses a liability sharing rule between the human user and the AI manufacturer. We assume that the policymaker is benevolent, in the sense that his objective function is aligned with that of society. In a second stage, the profit-maximizing AI manufacturer chooses the level of investment in AI autonomy and the price of the AI. In a third stage, the human user (who perfectly observes the AI's level of autonomy) decides whether or not to buy the AI and intervenes when the AI is not being used or is unable to act autonomously. The user's objective is to minimize her expected liability and intervention costs, as well as the price paid to the AI manufacturer. Ultimately, an accident may occur and payoffs are realized.

used without changing our results, as long as Assumption 1 holds.

[19] The decision rule $a^s(x;m)$ is exactly the one characterized in Section 2.2 of Gabaix (2019).

[20] To give an example, the function $h(a,x)=\frac{1}{2}(a-x)^2$ satisfies Assumption 1.

3.2 The first-best optimum

To conduct our normative analysis, we assume that the use of the AI is socially beneficial in that it avoids a costly human intervention. Specifically, human intervention is costly because (i) it requires a costly cognitive, and/or physical effort and (ii) the decision made by the user is less accurate than that made by the AI (to capture this idea, we assume that harm can only occur when the human user intervenes). Thus, at the first-best, all users should have access to the algorithm.

If everyone uses the AI, the expected social cost is:

$$SC(\pi) = (1-\pi)(k + H(m(\pi))) + c(\pi) \tag{5}$$

Assuming an interior solution, the first-best level of autonomy is characterized by the following first-order condition (FOC):

$$\frac{\partial SC}{\partial \pi}(\pi) = 0 \Leftrightarrow k - \frac{\mathrm{d}\left[(1-\pi)H(m(\pi))\right]}{\mathrm{d}\pi} = c'(\pi) \tag{6}$$

At the first-best level of autonomy, the marginal savings of increasing the AI autonomy, in terms of expected intervention costs and expected harm, should be equal to the marginal cost of that increase.

Proposition 1. *At the first-best optimum, all users have access to the algorithm, and the degree of AI autonomy is characterized by (6).*

3.3 The human user

Type of the human user. The human user is prone to behavioral inattention and can be either "naive" or "sophisticated".

A naive user is unaware that she is subject to behavioral inattention. Consequently, her expectation of harm in the event of an intervention is $H(1) = 0$. Thus, the only reason why a naive user might want to buy the AI is to reduce her cost of intervention (k) with probability π. She

will therefore choose to buy the AI if:

$$p \leq \pi k \equiv \underline{p}(\pi) \tag{7}$$

Since the naive user does not consider her attention limit, her willingness to pay $\underline{p}(\pi)$ ignores the fact that the AI outperforms her: the expected harm of an intervention does not effect her decision to use the AI.

A sophisticated user, on the other hand, is well aware of her limited level of attention. When using an AI, she expects that an intervention, which occurs with probability $(1-\pi)$, will cause an expected harm $H(m(\pi))$. She bears a fraction α of that expected harm. If she does not use an AI, she expects to carry the full expected harm $H(m(0))$ with certainty (she always intervenes and the manufacturer cannot be held liable for the harm, since the AI is not used). Note that, if she is subject to some degree of behavioral inattention, the expected harm in case of intervention is higher if she uses an AI ($H(m(\pi)) > H(m(0))$). Finally, the sophisticated user chooses to buy the AI if:

$$(1-\pi)[k + \alpha H(m(\pi))] + p \leq k + H(m(0)) \tag{8}$$

Which is equivalent to:

$$p \leq \pi\left[k + \alpha H(m(\pi))\right] + (1-\alpha)H(m(\pi)) - [H(m(\pi)) - H(m(0))] \equiv \overline{p}(\pi, \alpha) \tag{9}$$

On the one hand, the sophisticated user is willing to pay a higher amount to acquire the AI than the naive user, because she is aware that using the AI allows her to (i) avoid liability costs when the AI acts autonomously and (ii) benefit from shared liability in the event that she has to intervene (if $\alpha < 1$, the manufacturer bears a fraction of the harm). On the other hand, we see that the willingness to pay of the sophisticated user decreases with the difference $H(m(\pi)) - H(m(0))$. This is because, when the AI is used, the user's attention is lower, and therefore the expected harm in case of an intervention is higher. This effect is perfectly expected by sophisticated users (but not by naive users).

These two opposing effects make it possible that if the difference $H(m(\pi)) - H(m(0))$ is suffi-

ciently large, the willingness to pay of sophisticated users will be lower than that of naive users ($\overline{p}(\pi,\alpha) \leq \underline{p}(\pi)$).[21] It turns out that whether this is the case depends largely on how the degree of autonomy of the AI (π) affects the expected harm $(1-\pi)H(m(\pi))$. The expected harm decreases with the degree of autonomy if:

$$\frac{\mathrm{d}\left[(1-\pi)H(m(\pi))\right]}{\mathrm{d}\pi} < 0 \Leftrightarrow \frac{\mathrm{d}H\big(m(\pi)\big)}{\mathrm{d}\pi}\frac{1-\pi}{H\big(m(\pi)\big)} < 1 \tag{10}$$

The term on the left-hand side of the second condition in (10) can be interpreted as an elasticity: it is the percentage of increase in the expected harm $(1-\pi)H(m(\pi))$ when the probability of user intervention increases by 1% (in other words, when the AI's autonomy is reduced by 1%). From (10), the expected harm decreases with the level of autonomy of the AI if and only if the elasticity is less than 1.

More specifically, a higher π has two opposite effects on the expected harm. First, by reducing the probability of user intervention, a higher π reduces the probability that the user will cause harm: the expected harm decreases. Second, by reducing the user's level of attention, a higher π increases the harm in the event of an intervention, thus increasing the expected harm. The first effect dominates if condition (10) holds (the elasticity is less than 1), and conversely. Note that although we will focus on the case where the first effect dominates (Assumption 3), the second effect is likely to dominate in some specific cases, especially when the choice and/or the information structure is complex.[22]

Assumption 3. *When an algorithm is used, the expected harm* $(1-\pi)H(m(\pi))$ *decreases with the AI autonomy (i.e., condition (10) is satisfied).*

The next lemma follows from Assumption 3.

Lemma 1. *The willingness to pay of sophisticated users is higher than that of naive users* $(\overline{p}(\pi,\alpha) \geq \underline{p}(\pi)\forall(\pi,\alpha))$.

[21] Note that if the difference $H(m(\pi)) - H(m(0))$ is very large, it is also possible that using the AI is no longer socially beneficial.

[22] Indeed, some authors have shown that behavioral biases can undermine the efficiency of combining the information from an AI and a human user. For example, Agarwal et al. (2023) show that users can make belief updating errors and fail to correctly account for the correlation between their own information (obtained through their expertise) and the information obtained through the AI prediction task. As a result, the situation where a human is assisted by an AI may be suboptimal compared to cases where only the human or the AI information is used.

Proof. If $\pi = 0$, the willingness to pay of the sophisticated user is $\overline{p}(0, \alpha) = (1-\alpha)H(m(0)) \geq 0$ and that of the naive user is $\underline{p}(0) = 0$. Thus, if $\pi = 0$, the willingness to pay of the sophisticated user is higher than that of the naive user:

$$\overline{p}(0, \alpha) \geq \underline{p}(0) \tag{11}$$

For all $\pi \geq 0$, we can rewrite the willingness to pay of the sophisticated user:

$$\overline{p}(\pi, \alpha) = \underline{p}(\pi) - \alpha(1 - \pi)H(m(\pi)) + H(m(0)) \tag{12}$$

From which:

$$\frac{\partial \overline{p}}{\partial \pi}(\pi, \alpha) = \underline{p}'(\pi) - \alpha \frac{\mathrm{d}\left[(1 - \pi)H(m(\pi))\right]}{\mathrm{d}\pi} \tag{13}$$

With $\underline{p}'(\pi) = k > 0$. According to Assumption 3, it follows from (10) and (13) that:

$$\frac{\partial \overline{p}}{\partial \pi}(\pi, \alpha) \geq \underline{p}'(\pi) \tag{14}$$

From (11) and (14), and since $\underline{p}$ and $\overline{p}$ are continuous with respect to π, we can deduce that $\overline{p}(\pi, \alpha) \geq \underline{p}(\pi) \forall (\pi, \alpha)$. □

In the following, we denote by β the fraction of human users who are naive, and $1 - \beta$ the fraction of users who are sophisticated.

3.4 The manufacturer

The AI manufacturer has to make two decisions: (i) the price at which the AI is sold, and (ii) the level of autonomy of the AI.

The manufacturer is assumed to be a monopolist.[23] Consequently, if he decides to develop and market the AI (*i.e.* he sets a price p low enough that the AI will be bought by at least some

[23] We can expect that the assumption of perfect competition in the market for AI will significantly alter some of our results in the following ways. AI manufacturers will differentiate their algorithms by offering specific levels of autonomy tailored to either naive users or sophisticated users, depending on their preferences. Since the willingness to pay of sophisticated users increases more rapidly with AI autonomy than that of naive users (see the proof of Lemma 1), the AI offered to sophisticated users will have more autonomy and will be more expensive. The level of autonomy chosen by manufacturers developing AI for naive users is too low compared to the first best, while it is socially optimal for AI intended to be used by sophisticated users.

users), he will have to choose between (i) a price $\underline{p}(\pi)$ that all users, regardless of their type, will pay (sophisticated users get a surplus) and (ii) a higher price $\overline{p}(\pi, \alpha)$ that extracts all the surplus from sophisticated users, but for which naive users are not willing to buy the AI (only sophisticated users are willing to pay this price). We assume that the manufacturer also has the option not to distribute the AI, in which case he saves the fixed cost $c(0)$ by not investing in its development, and receives no profit.

Let us first assume that the AI manufacturer sets the price $\underline{p}(\pi)$ (the AI is sold to all users, regardless of their type). In this case, his expected payoff is:

$$\Pi_{\underline{p}}(\pi, \alpha) = \pi k - (1 - \pi)(1 - \alpha)H(m(\pi)) - c(\pi) \tag{15}$$

The FOC for the choice of the degree of autonomy of the AI (π) is:

$$\frac{\partial \Pi_{\underline{p}}}{\partial \pi}(\pi, \alpha) = 0 \Leftrightarrow k - (1 - \alpha)\frac{\mathrm{d}\left[(1 - \pi)H(m(\pi))\right]}{\mathrm{d}\pi} = c'(\pi) \tag{16}$$

We denote by $\pi^*_{\underline{p}}(\alpha)$ the degree of autonomy implicitly defined by this FOC. For the manufacturer, there are two marginal benefits to increasing π. First, increasing π reduces the probability that an intervention will be required, and thus the expected cost of an intervention to a user. As a result, the value of the AI to a user increases, and the manufacturer is able to sell the AI at a higher price (the first term on the left-hand side of (16)). Second, according to Assumption 3, an increase in π reduces the expected harm, and thus the manufacturer's expected liability cost (the second term on the left-hand side of (16)).

Lemma 2. *The expected profit of the AI manufacturer, given that he chooses a price* $\underline{p}(\pi^*_{\underline{p}}(\alpha))$*, decreases with his share of liability* $(1 - \alpha)$*.*

Proof. Using the envelope theorem, we find that Assumption 3 implies that $\frac{\partial \Pi_{\underline{p}}}{\partial \alpha}(\pi^*_{\underline{p}}(\alpha), \alpha) \geq 0$. □

Now let us assume that the AI manufacturer sets the price $\overline{p}(\pi, \alpha)$ (the AI is sold only to

sophisticated users). In this case, the AI manufacturer's expected profit, after simplification, is:

$$\Pi_{\underline{p}}(\pi) = (1-\beta)\left[\pi(k + H(m(\pi))) - (H(m(\pi)) - H(m(0)))\right] - c(\pi) \tag{17}$$

Note that this expected profit does not depend on the liability sharing rule (α). The FOC for choosing the degree of autonomy of the AI is:

$$\frac{\partial \Pi_{\underline{p}}}{\partial \pi}(\pi) = 0 \Leftrightarrow (1-\beta)\left[k - \frac{\mathrm{d}\left[(1-\pi)H(m(\pi))\right]}{\mathrm{d}\pi}\right] = c'(\pi) \tag{18}$$

We denote by $\pi^*_{\underline{p}}$ the degree of autonomy implicitly defined by this FOC. In contrast to the case where the price is $\underline{p}(\pi)$, increasing π now reduces the manufacturer's share of the damage, either indirectly through the price (for a fraction α of the damage) or directly through the liability sharing rule (for a fraction $1-\alpha$ of the damage). As a result, the liability sharing rule does not affect the choice of π in this case. Another difference is that the marginal revenue of the manufacturer is discounted by the fraction of sophisticated users $(1-\beta)$, since naive users do not buy the AI.

What is the choice of the AI manufacturer between the "low" price $\underline{p}(\pi)$ and the "high" price $\overline{p}(\pi, \alpha)$? This choice depends on the proportion of naive versus sophisticated users. To understand why, let us first assume that $\beta = 0$ (*i.e.*, all users are sophisticated). In this case, we have $\Pi_{\underline{p}}(\pi) > \Pi_{\underline{p}}(\pi, \alpha)$ for all levels of autonomy π, and the AI manufacturer is better off choosing a price $\overline{p}(\pi^*_{\underline{p}}, \alpha)$ that allows him to obtain an expected profit $\Pi_{\underline{p}}(\pi^*_{\underline{p}}) > 0$. Now let us assume that $\beta = 1$ (*i.e.*, all users are naive). In this case, we have $\Pi_{\underline{p}}(\pi) \leq 0$, and the manufacturer may prefer to sell at price $\underline{p}(\pi^*_{\underline{p}}(\alpha))$.

However, if the proportion of naive users is high (*e.g.*, $\beta = 1$), the manufacturer may not be able to make a positive expected profit. Together with Lemma 2, this implies that if the proportion of naive users is high and the AI manufacturer is liable for a significant fraction of the expected harm (α is low), he may choose not to develop the AI, even though this choice increases the expected cost of human intervention and the expected harm (recall that, by assumption, developing and distributing the AI to all users is socially beneficial).

3.5 The socially optimal sharing of liability

The expected social cost is function of whether all users or only sophisticated users buy the AI, which in turn depends on the price level.

3.5.1 Optimal sharing rule for given prices

Suppose the AI manufacturer chooses a "low" price $\underline{p}(\pi^*_{\underline{p}}(\alpha))$ such that all users buy the AI. A comparison using the FOCs (6) and (16) shows that $\pi^*_{\underline{p}}(\alpha)$ is inferior to the first-best, unless the AI manufacturer is fully liable for the damage ($\alpha = 0$).[24] Otherwise (if $\alpha > 0$), the expected liability cost faced by the manufacturer does not allow him to fully internalize the marginal benefit, in terms of reduced expected harm, of increasing the autonomy of the AI. As a result, the manufacturer underinvests in autonomy.

Note that if $m(\pi) = 1$ (no behavioral inattention, regardless of the level of AI autonomy), users will always choose the action that is objectively the most appropriate ($a = x$). There is no expected harm, and thus the liability rule is inconsequential: the AI manufacturer's objective is then always aligned with that of society.[25]

Now suppose that the AI manufacturer chooses a "high" price $\overline{p}(\pi^*_{\underline{p}}, \alpha)$ so that only sophisticated users will buy the AI. How does the manufacturer's choice compare to the first-best? On the one hand, excluding naive users from using the AI is socially costly, both because naive users cause harm that could be avoided by using the AI, and because the marginal social gain is now discounted, resulting in a lower level of AI autonomy. On the other hand, given the fact that only sophisticated users buy the AI, it is possible to show that the manufacturer's choice of AI autonomy is socially optimal.

On the latter point, since only sophisticated users buy the AI, the expected social cost is:

$$SC_+(\pi) = (1-\beta)(1-\pi)(k + H(m(\pi))) + \beta(k + H(m(0))) + c(\pi) \tag{19}$$

[24] Note that when $\alpha = 0$, the user does not necessarily choose the most appropriate action according to his subjective perception ($a = a^s(x^s(m); m)$). However, since the user is then indifferent between the action $a^s(x; m)$ and any other action (he internalizes no harm), we assume that he chooses $a = a^s(x; m)$.

[25] If $m = 1$, both types of users get the same payoffs (hence they behave in the same way), and the prices $\underline{p}(\pi)$ and $\overline{p}(\pi, \alpha)$ converge: $\underline{p}(\pi) = \overline{p}(\pi, \alpha) = \pi k$.

For a "high" price $\overline{p}(\pi^*_{\underline{p}}, \alpha)$, the manufacturer chooses an AI autonomy level π^*_2, and the expected social cost is $SC_+(\pi^*_2)$. The AI autonomy level and the expected social cost are independent of the share of liability α. Moreover, the autonomy level that minimizes the expected social cost is characterized by a FOC equivalent to (18), which characterizes the autonomy level $\pi^*_{\underline{p}}$, implying that the AI autonomy chosen by the manufacturer minimizes the expected social cost.

The following proposition summarizes the previous findings.

Proposition 2. *If the manufacturer chooses the "low" price $\underline{p}(\pi^*_{\underline{p}}(\alpha))$, the AI autonomy is socially optimal only if $\alpha = 0$. If the manufacturer chooses the "high" price $\overline{p}(\pi^*_{\underline{p}}, \alpha)$, the AI autonomy is socially optimal (given that only sophisticated users use the AI) regardless of the liability sharing rule.*

Note also, with respect to AI diffusion, that $SC(\pi) > SC_+(\pi) \forall \pi$, confirming that excluding naive users is socially costly.

3.5.2 Optimal sharing rule with endogenous pricing

So far, we have focused on the optimal liability sharing rule separately for prices $\underline{p}(\pi^*_{\underline{p}}(\alpha))$ and $\overline{p}(\pi^*_2, \alpha)$. Since, as explained above, it is preferable not to exclude users from using the AI, and since some users will indeed be excluded if the manufacturer sets a high price, we are interested in the effect of the liability sharing rule (α) on the manufacturer's pricing decision. Using the envelope theorem, we have:

$$\frac{\mathrm{d}\Pi_{\underline{p}}(\pi^*_{\underline{p}}(\alpha), \alpha)}{\mathrm{d}\alpha} = (1 - \pi^*_{\underline{p}}(\alpha)) H(m(\pi^*_{\underline{p}}(\alpha))) > \frac{\mathrm{d}\Pi_{\underline{p}}(\pi^*_{\underline{p}})}{\mathrm{d}\alpha} = 0 \tag{20}$$

Thus, as the share of liability borne by the user (α) increases, it becomes relatively more profitable for the manufacturer to charge a "low" price $\underline{p}(\pi^*_{\underline{p}}(\alpha))$.

Proposition 3. *Increasing the user's share of liability (α) may induce the manufacturer to lower its price.*[26]

Thus, increasing the user's share of liability (α) may be socially beneficial in that naive users

[26] More specifically, increasing α from a level α' to a level α'' improves the diffusion of the AI by inducing naive users to buy it when $\Pi_{\underline{p}}(\pi^*_{\underline{p}}(\alpha'), \alpha') < \Pi_{\underline{p}}(\pi^*_{\underline{p}}) < \Pi_{\underline{p}}(\pi^*_{\underline{p}}(\alpha''), \alpha'')$.

will then use the AI, reducing the total expected cost through both the expected cost of user intervention and the expected harm.

In summary, we find that because some users may be prone to some degree of behavioral inattention without being aware of it, the policymaker may face a trade-off when setting the liability sharing rule.[27] On the one hand, if the price is low, increasing the liability of the AI manufacturer brings his objective closer to that of society and incentivizes him to choose a higher level of AI autonomy. On the other hand, reducing the liability of the manufacturer has two distinct advantages. First, it helps to avoid situations in which the AI manufacturer does not develop or sell the AI at all. Second, if the AI is indeed developed and sold, the AI manufacturer may respond by lowering its price, allowing a larger fraction of users to benefit from the AI.

4 The cost of attention

In the extension presented in this section, we assume that the user's level of attention m is a choice variable that can be increased by costly effort (the choice is made before the user may have to intervene).[28] We now assume that all users are sophisticated ($\beta = 0$). The main reason for this assumption is that it is not possible to model the effort choice of a naive user, since by definition she is not aware that her attention is imperfect and therefore cannot be aware of the possibility of improving that effort.

4.1 The human user

Let us assume that the cost of the user's attentional effort is an increasing, convex function of his attentional level, with $c_o'(0) = 0$, $c_o'(m) > 0$ and $c_o''(m) > 0$. For simplicity, we also assume that the cost of intervention is zero ($k = 0$).

If the user chooses to use the AI, she faces the following expected cost:

$$(1 - \pi)\alpha H(m) + c_o(m) + p \tag{21}$$

[27] In the absence of behavioral inattention, *i.e.*, if $m(\pi) = 1 \forall \pi$, the first-best is always achieved: all users buy the AI, and the AI autonomy chosen by the manufacturer minimizes the expected social cost.

[28] The idea that attention (and thus decision making) can be improved by costly cognitive effort already exists in the literature on rational inattention (see, *e.g.*, Sims, 2003; Caplin and Dean, 2015).

The FOC for the level of attention is:

$$-(1-\pi)\alpha H'(m) = c_o'(m) \tag{22}$$

We denote by $\underline{m}(\pi, \alpha)$ the level of attention of the user characterized by this FOC. Increasing her level of attention is (cognitively) costly (right-hand side of (22)), but it reduces the expected harm, and thus her expected liability cost (left-hand side of (22)). This marginal benefit, and thus the user's level of attention, decreases with the AI's degree of autonomy (π) and increases with the user's liability share (α):[29]

$$\frac{\partial \underline{m}}{\partial \pi}(\pi, \alpha) \leq 0 \text{ and } \frac{\partial \underline{m}}{\partial \alpha}(\pi, \alpha) > 0 \tag{23}$$

We also have:

$$\frac{\partial^2 \underline{m}}{\partial \pi \partial \alpha}(\pi, \alpha) < 0 \tag{24}$$

The negative sign of this cross-derivative means that while the user's attention increases with her share of liability, this increase in attention is smaller the greater the degree of autonomy of the AI.[30]

If she chooses not to use the AI, the user faces the expected cost:

$$H(m) + c_o(m) \tag{25}$$

The FOC is:

$$H'(m) = c_o'(m) \tag{26}$$

The level of attention of the user characterized by this FOC is denoted by $\overline{m}$. Since the manufacturer cannot be held liable when the AI is not in use, the level of attention $\overline{m}$ is independent of the liability sharing rule. Note also that $\overline{m} \geq \underline{m}(\pi, \alpha)$: the level of attention chosen by the user is higher when she is not using the AI. This is consistent with the assumptions

[29] As in the baseline model of the previous section, we find that increasing the autonomy of the AI leads to a reduced level of attention on the part of the user, except that this effect is now endogenous to the model. This effect exists only if the manufacturer is liable for some portion of the expected harm, *i.e.*, if $\alpha < 1$.

[30] Another possible interpretation is that the user's attention decreases with to the degree of autonomy of the AI, but an increase in the user's liability reduces this loss of attention.

made in the previous section, with the difference that the user's level of attention when using an AI is now increases with the user's liability.

Finally, the user will acquire the AI if:

$$p \leq [c_o(\overline{m}) - c_o(\underline{m}(\pi, \alpha))] + [H(\overline{m}) - (1-\pi)\alpha H(\underline{m}(\pi, \alpha))] \equiv p^*(\pi, \alpha) \tag{27}$$

The price $p^*(\pi, \alpha)$ that the user is willing to pay for the AI is the sum of the savings in (i) the user's attention cost and (ii) the user's expected liability cost.

4.2 The manufacturer

After simplification, the manufacturer's expected profit from developing and marketing the AI is given by:

$$\Pi(\pi, \alpha) = [c_o(\overline{m}) - c_o(\underline{m}(\pi, \alpha))] + [H(\overline{m}) - (1-\pi)H(\underline{m}(\pi, \alpha))] - c(\pi) \tag{28}$$

The manufacturer's expected profit is equal to (i) the reduction in the cost of attention, plus (ii) the reduction in the expected harm, minus (iii) the investment in the AI's autonomy ($c(\pi)$). Note that the expected harm is fully internalized by the manufacturer (*via* the price and the manufacturer's liability).

The FOC for the autonomy level of the AI is:

$$\frac{\partial \Pi}{\partial \pi}(\pi, \alpha) = H(\underline{m}(\pi, \alpha)) - c'(\pi) - \frac{\partial \underline{m}}{\partial \pi}(\pi, \alpha)[(1-\pi)H'(\underline{m}(\pi, \alpha)) + c_o'(\underline{m}(\pi, \alpha))] = 0 \tag{29}$$

Substituting the user's FOC (22) into (29), we have:

$$\frac{\partial \Pi}{\partial \pi}(\pi, \alpha) = H(\underline{m}(\pi, \alpha)) - c'(\pi) - \frac{\partial \underline{m}}{\partial \pi}(\pi, \alpha)(1-\pi)(1-\alpha)H'(\underline{m}(\pi, \alpha)) = 0 \tag{30}$$

We denote by $\pi^*(\alpha)$ the AI autonomy chosen by the manufacturer characterized by (30). Note that the level of autonomy chosen by the manufacturer now depends on the liability sharing rule. In the following, in order to focus on the trade-off faced by the policymaker, we assume that $\Pi(\pi^*(\alpha), \alpha) > 0 \forall \alpha$ (*i.e.*, the manufacturer is always willing to develop and market the AI),

which implies that $p^*(\pi^*(\alpha), \alpha) > 0 \forall \alpha$.

4.3 The socially optimal sharing of liability

The expected social cost is:

$$SC_*(\alpha) = (1 - \pi^*(\alpha))H(\underline{m}(\pi^*(\alpha), \alpha) + c_o(\underline{m}(\pi^*(\alpha), \alpha) + c(\pi^*(\alpha)) \tag{31}$$

Using (30) and (22), the FOC for the optimal liability sharing rule is:

$$\frac{\partial SC}{\partial \alpha}(\alpha) = \frac{\partial \underline{m}}{\alpha}(\pi^*(\alpha), \alpha)(1 - \pi^*(\alpha))(1 - \alpha)H'(\underline{m}(\pi^*(\alpha), \alpha)) = 0$$

A no liability rule for the manufacturer ($\alpha = 1$) satisfies this FOC.

Proposition 4. *Assume that the user can increase his level of attention at a cost and that all users are sophisticated ($\beta = 0$). No liability of the manufacturer ($\alpha = 1$) is socially optimal.*

The intuition is as follows. Compared to the baseline model, given that the user is always of the sophisticated type, increasing the manufacturer's share of liability is not optimal, since the manufacturer already fully internalizes the expected harm (via the price and/or the liability rule). In fact, in the baseline model, only the presence of naive users means that the harm is not fully internalized by the manufacturer, who is consequently able to extract an additional rent by reducing his investment in the autonomy of the AI. This effect disappears when $\beta = 0$ (as we have assumed in this section). On the other hand, increasing the user's share of liability increases her level of attention, which is suboptimal if she does not fully internalize the harm (*i.e.*, if she is not fully liable for the expected harm). In a more comprehensive model, with both sophisticated and naive users, as well as the possibility of costly attentional effort, we can expect a more nuanced trade-off between (i) increasing the manufacturer's share of liability to improve the AI's degree of autonomy, and (ii) increasing the user's share of liability to improve both the AI's diffusion and the user's attentional effort.

5 Conclusion

In this paper, we highlight some of the trade-offs involved in choosing the socially optimal liability sharing rule between the manufacturer of a performative artificial intelligence (AI) algorithm and the human user of that AI. To this end, we propose a model in which we assume that even when using an AI, the human user must intervene when the AI fails to handle a situation. More situations can be handled by the AI as its autonomy increases (through costly investments by the AI manufacturer). We have also assumed that the performance of the AI is better than that of the human user. In fact, the latter may be subject to behavioral inattention (Gabaix, 2019). This inattention, which is expected to increase with the level of autonomy of the AI, can lead to poor decisions that can cause harm. To limit the expected harm and the cost of user intervention, it is therefore important both to encourage the manufacturer to invest sufficiently in the autonomy of the AI, and to ensure that as many human users as possible have access to the AI. Another important assumption of our model is that only a fraction of the users (the "sophisticated" users) are aware of their attentional limitations, while the rest of the users (the "naive" users) do not consider the cost of their inattention when not using an AI or when confronted with a situation that the AI cannot handle.

Our results show that the problem of inattention, coupled with the lack of awareness of some users of their attentional limitations, can lead to insufficient diffusion of an AI, resulting in a loss of social welfare. Policymakers can limit this problem by reducing the share of liability borne by the AI manufacturer in the event of harm. Another benefit of reducing the manufacturer's liability (and thus increasing the user's liability) is that users will have an incentive to be more attentive (which is beneficial if they have control over their level of attention). However, we show that the AI autonomy chosen by the manufacturer tends to be too low, and that reducing his liability would exacerbate the problem. Thus, when choosing the liability sharing rule, policymakers face a trade-off between (i) supporting the diffusion of AI and increasing the attention level of the users, and (ii) incentivizing the manufacturer to invest more in the AI autonomy.

Some of the assumptions in our model are worth discussing. First, we did not consider the possibility that the AI user could modulate the frequency of AI use. However, the emerging

literature on autonomous vehicles has highlighted the importance of considering the user's choice of activity level (see, *e.g.*, Shavell, 2020). Second, we have not considered other possible liability regimes, such as the negligence rule. However, the context in which we find ourselves makes it difficult to use such a rule, because the standard would then have to refer to a minimum level of autonomy to be achieved, which raises problems that are difficult to solve in terms of the concrete formulation of the standard and the incentive to innovate (see Dawid and Muehlheusser, 2022). Other reasons why the negligence rule seems difficult to apply in a context where harm is potentially caused by the failure of an AI are discussed in Obidzinski and Oytana (2022). Finally, to the extent that the manufacturer is able to sell the use of the AI he has developed at a higher price to sophisticated users, it might be interesting to let him educate consumers in order to increase the proportion of sophisticated users relative to the proportion of naive users. This possibility could have some additional social benefits.

Despite the limitations of our assumptions, we believe that our approach, which introduces behavioral inattention into a liability model, highlights new trade-offs that complement those already existing in the literature on liability rules for defective products and, more specifically, performative algorithms such as autonomous vehicles.

References

Agarwal, N., Moehring, A., Rajpurkar, P., and Salz, T. (2023). Combining human expertise with artificial intelligence: Experimental evidence from radiology. NBER Working paper.

Alberdi, E., Strigini, L., Povyakalo, A. A., and Ayton, P. (2009). Why are people's decisions sometimes worse with computer support? In Computer Safety, Reliability, and Security: 28th International Conference, SAFECOMP 2009, Hamburg, Germany, September 15-18, 2009. Proceedings 28, pages 18–31. Springer.

Armstrong, M. and Vickers, J. (2012). Consumer protection and contingent charges. Journal of Economic Literature, 50(2):477–493.

Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6):775–779.

Caplin, A. and Dean, M. (2015). Revealed preference, rational inattention, and costly information acquisition. American Economic Review, 105(7):2183–2203.

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

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.

Gabaix, X. (2019). Behavioral inattention. In Handbook of behavioral economics: Applications and foundations 1, volume 2, pages 261–343. Elsevier.

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

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

Obidzinski, M. and Oytana, Y. (2022). Prediction, human decision and liability rules. CRED Working paper No 2022-06.

O'Donoghue, T. and Rabin, M. (1999). Doing it now or later. The American Economic Review, 89(1):103–124.

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

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

Sims, C. A. (2003). Implications of rational inattention. Journal of Monetary Economics, 50(3):665–690.

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

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