Artificial intelligence, inattention and liability rules

Marie Obidzinski, Yves Oytana

CRED
Centre de recherche
en économie
et droit

CRED WORKING PAPER n^o 2024-07

Artificial intelligence, inattention and liability rules

June, 2024

MARIE OBIDZINSKI* YVES OYTANA†

*Université Paris-Panthéon-Assas, CRED, F-75005, Paris, France.
†Université de Franche-Comté, CRESE, F-25000 Besançon, France.

Artificial intelligence, inattention and liability rules

Marie Obidzinski,* Yves Oytana†

May 31, 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 assumptions relate 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, F-25000 Besançon, France. e-mail: yves.oytana@univ-fcomte.fr

1 Introduction

It is well known that driver inattention is a major cause of traffic accidents. According to Knowles and Tay (2002), "the relative contribution of driver inattention to the social cost of road crashes is expected to be much higher compared to drink driving or speeding".[1] The advent of autonomous vehicles (AVs) is expected to circumvent this inattention problem, and to significantly reduce the risk of accidents.[2] However, at the present time, fully autonomous AVs have not yet been developed. The Society of Automotive Engineers has developed a six-level classification system, ranging from no automation (Level 0) to full automation (Level 5). Partial automation (Level 2) requires the driver to keep his or her hands on the wheel.[3] Conditional automation (Level 3) covers all aspects of driving, but the human driver must be ready to intervene. Such high-tech cars with conditional automation have recently been introduced to the market.[4]

Partial and conditional automation means that even with so-called AVs, the driver should remain alert. In fact, the problem of inattention may even be exacerbated: the difficulty of maintaining attention when using autonomous artificial intelligence (AI) has long been recognized in computer science (see, *e.g.*, Parasuraman and Riley, 1997; Zerilli et al., 2019). According to Bainbridge (1983), "it is impossible for even a highly motivated human being to maintain effective visual attention towards a source of information on which very little happens, for more than about half an hour." One explanation for this observation is that automation can lead to "passive fatigue" (Desmond and Hancock, 2001; Saxby et al., 2013). Passive fatigue is caused by reduced task engagement and typically occurs during monotonous monitoring tasks that require little operator intervention. Thus, when faced with a situation that an AV cannot handle, the driver disengagement can lead to increased crash risk.

[1] Moreover, as shown by Klauer et al. (2006), the risk of crashes increases significantly as drivers engage in various types of inattention-related activities. Along the same lines, Cunningham and Regan (2018) emphasize that "driver inattention including distraction is purported to be a contributing factor in around 65% of safety-critical events."

[2] The idea that safer AVs can mitigate the risks posed by driver inattention is acknowledged, for example, by Dawid et al. (2024): "a safer AV can prevent some accident which might have been caused by the [human-driven vehicle (HV)], *e.g.* due to inattention or careless behavior of the HV's driver."

[3] Advanced Driver Assistance Systems (ADAS) do not cover all aspects of driving.

[4] Source: The Wall Street Journal, September 28, 2023, https://www.wsj.com/lifestyle/cars/mercedes-benz-hands-free-drive-pilot-review-tesla-comparison-ede47ac7.

Passive fatigue is not limited to AV monitoring, but extends to the use of other autonomous AI.[5] In aeronautics, for example, projects are being developed to increase the autonomy of aircraft.[6] However, these systems still require the operator (*i.e.*, the pilot) to intervene occasionally. Inattention problems also arise in the use of decision support algorithms. 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.[7]

The consequences of inattention can be all the more serious because consumers are not always aware of the risks associated with distraction and/or have a biased perception of the quality of their attention. Individuals are often unaware of how distracted they may be, as evidenced by their lack of awareness of digital media consumption. This is shown, for example, by Andrews et al. (2015) in the context of an experiment comparing reported and actual phone use. They found that the number of phone uses was significantly underreported, while there was only a moderate positive correlation between reported and actual duration of phone use. More recently, Parry et al. (2021) examined the relationship between self-reported and actual digital media use as part of a meta-analysis. Their main conclusion was that "self-reports were rarely an accurate reflection of logged media use." These findings suggest that drivers (and, more generally, users of semi-autonomous AI and advisory algorithms) may not be realistic about the detrimental effects of their inattention and the extent of their distraction.

But inattention problems in the use of autonomous AI and advisory algorithms do not mean that these technologies should not be used. On the contrary, their use generally reduces the overall risk

[5] Autonomous AI is defined as a type of AI that uses the latest advances in technology to enable devices to perform tasks independently of the human operator.

[6] See, for example, the DragonFly project: https://projectdragonfly.nl/.

[7] In a companion paper (Obidzinski and Oytana, 2022), we focus on the automation bias in the case of advisory algorithms, *i.e.* the tendency to place more weight or trust in our own human capabilities. 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 Alberdi et al., 2009; Cummings, 2017; Parasuraman and Riley, 1997).

of accidents.[8] However, it is important to find ways to mitigate the risks specifically associated with inattention. Various approaches have been proposed in the literature, such as designing devices to attract the operator's attention, or conceiving systems to ensure a minimum level of active operator involvement. In this paper, we focus on a different instrument, namely liability. Indeed, an efficient allocation of liability between an AI manufacturer and the user of that AI can help mitigate the welfare costs of inattention. Specifically, this paper examines how liability affects the manufacturer's investment in improving AI autonomy, the user's level of attention when using the AI, and the diffusion of the AI.

To determine how best to allocate liability, we develop a game-theoretic framework in which three players interact: a policymaker, an AI manufacturer, and a human user of the AI. First, the policymaker chooses a liability sharing rule to allocate damages between the AI manufacturer and the user in the event of an accident while using the AI. Next, the AI manufacturer chooses both the degree of autonomy of the AI (through a costly investment) and a selling price. Finally, the human user decides whether or not to buy the AI system. If she decides not to buy, her intervention is always necessary. Conversely, if she decides to buy the AI, her intervention is only necessary if a situation arises that the AI cannot handle (the likelihood of such a situation occurring decreases as the AI becomes more autonomous). When the human user is required to intervene, she chooses an action that may cause some harm if it is not appropriate. The appropriateness of her action depends on her level of attention.

The two main assumptions of our model are related to attention. First, we assume that human users are inattentive in the sense of Gabaix (2019). The framework provided by Gabaix (2019) has the advantage of providing a tractable way to model different degrees of behavioral inattention. Second, we assume that at least some users are not fully aware of the detrimental consequences of their inattention and/or misperceive their attentional capacity. Formally, it is only a subset of the users (the "sophisticated" users) who have a correct perception of the damage that is caused by their inattention. The remaining users (the "naive" users) underestimate the harm resulting from

[8] However, the nature of the risk associated with the use of AI is often different. For example, an autopilot may be the cause of an accident that would have been easily avoided by a human driver, and vice versa.

their inattention and/or overestimate their ability to remain attentive. An additional assumption of our model is that the use of the AI by all users is socially beneficial because (i) the AI performs better than the human (it reduces the expected harm by taking more appropriate actions), and (ii) it reduces the probability of a costly human intervention (an intervention is costly to the user because it involves cognitive and/or physical effort).

In this setting, we show that the AI manufacturer's investment in improving the AI autonomy is insufficient if users are subject to behavioral inattention. The policymaker can incentivize the AI manufacturer to invest more by increasing his liability. However, this should be balanced against two types of costs. First, increased manufacturer liability may slow down the diffusion of AI (AI will only be purchased by sophisticated users). Second, in more extreme cases, the AI manufacturer may even stop developing and marketing the AI system. As an extension to our baseline model, we assume that the user can manage higher levels of attention through a costly cognitive effort. We show that the chosen cognitive effort increases with user liability, adding a new element to the previous trade-off.

The remainder of the paper is organized as follows. Section 2 discusses the related literature. In Section 3, we present a simple example to illustrate the trade-off the policymaker faces when choosing the liability rule. Section 4 lays out the baseline model and the optimal liability sharing rule is characterized in Section 5. In Section 6, we present some extensions to our baseline model. Section 7 provides further discussion and concludes.

2 Related literature

Our article is related to the literature on product liability (see, *e.g.*, Landes and Posner, 1985; Daughety and Reinganum, 2013; Hay and Spier, 2005). Specifically, Hay and Spier (2005) propose a bilateral care model in which consumers may harm a third party when using a product. When consumers have deep pockets (their wealth covers the damage), 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 consumers' wealth is so low that they are unable to pay for the entire damage. In this case, liability sharing between the user and the manufacturer is a second-best solution.[9] In our model, liability sharing is also a second-best solution, although it is explained by a trade-off between AI diffusion and mitigating attention problems.

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.

Several authors have focused on liability rules in the specific case of autonomous vehicles (see, *e.g.*, Shavell, 2020; Di et al., 2020). In this strand of the literature, the contributions that are most similar to ours are those that simultaneously consider the investment made in the development phase of AVs and the decision of individual drivers whether or not to purchase an AV. This is the case of De Chiara et al. (2021). They show that a strict liability rule is socially preferable to a negligence rule, because the producer does not internalize the expected damage of accidents under the negligence rule. In contrast to our setting, De Chiara et al. (2021) do not consider the possibility of liability sharing between the producer and the user of the AV, nor do they focus on attention issues. Dawid et al. (2024) consider a mixed traffic model (with both human-driven vehicles and AVs) in which the demand is endogenous. In their model, the demand for both types of vehicles depends on the safety investments made by the AV producer, which directly benefit AV users, but also indirectly benefit the human-driven vehicle (HV) by making the road safer. They also consider a new policy instrument: the level of V2I connectivity infrastructure, which allows a better connectivity between AVs and thus reduces the risk of collisions between AVs. Notably,

[9] In particular, a "residual manufacturer liability" (the manufacturer has to pay the shortfall of damages not paid by the consumer) limits the inefficiencies arising from the judgment proof problem.

Dawid et al. (2024) find a trade-off close to the one we highlight in our paper (although the rationale is different from ours): increased producer liability incentivizes him to invest more in safety, but reduces AV market penetration. The paper by Feess and Muehlheusser (2024), while not directly dealing with liability, has some interesting similarities to our work. They consider the protection of passengers relative to third parties (*e.g.*, pedestrians) in the design of AVs. According to their results, the second-best level of passenger protection should be higher than the first-best level in order to increase the market penetration of AVs (drivers are heterogeneous with respect to their costs of AV adoption) and the level of care of third parties.

3 Liability, AI 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: AI autonomy, expected harm and development cost.

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: AI autonomy and expected social cost.

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,[10] or at a "high" price, so that only sophisticated users buy it.[11] 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.[12] If we switch to a strict liability rule (the manufacturer must pay for all the harm), table 3 now shows that

[10]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).

[11]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.

[12]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).

the manufacturer receives the highest expected profit by choosing a high level of AI autonomy.[13] 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: Manufacturer's expected profits.

Under strict liability, however, the manufacturer may be tempted to deviate by choosing a higher price (to reduce his expected liability). 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:[14] 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 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.[15]

4 The Model

We formalize the trade-off highlighted in the previous section by building a game-theoretic framework with three players: a policymaker, an AI manufacturer (he), and a human user (she). In this section, we first present the general setup of our model by describing the players' objectives, their

[13] 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).

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

[15] 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.

decision variables, and the way we formalize the behavioral assumptions related to attention. This is followed by a description of the timing of the game.

The policymaker. The policymaker's objective is to minimize the expected social cost by choosing a liability rule. The liability rule is represented by the fraction $\gamma \in [0, 1]$ that the human user bears in case of an accident while using the AI. The AI manufacturer is liable for the remaining part $(1 - \gamma)$.

The AI manufacturer. The AI manufacturer is assumed to be a monopolist. He chooses the price p and the level of autonomy $\pi \in [0, 1)$ of the AI that maximizes his expected profit (*i.e.*, his expected revenue, minus expected liability costs and the cost of investing in the AI's autonomy), given the liability sharing rule γ. With probability π, the AI will not request human intervention. However, with the complementary probability $1-\pi$, the AI will encounter a situation that it cannot handle autonomously, and human intervention is required. For example, consider self-driving cars with conditional automation. The driver may need to intervene in time to respond appropriately to a request from the car's AI. It is assumed that the level of autonomy of the AI is observable to the human user when deciding whether to purchase the AI. 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 $c(0) > 0$. As explained below, users' willingness to pay depends on their level of sophistication. Therefore, by choosing the price, the manufacturer may decide to satisfy the demand of only some types of users. 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.

The human user. The human user of the AI decides whether to buy the AI and, if the AI requests human intervention, intervenes by choosing an action a. The user's objective is to minimize her expected liability and intervention costs, as well as the price she pays for the AI. Central to our analysis is the assumption that the human user may be inattentive in the sense of Gabaix (2019), and that she is unaware of her behavioral inattention if she is of the naive type (with an exogenous probability β).

Inattention. Inattention is represented as follows. In the case of an intervention, the user knows that the appropriate action is the realization x of a random variable X, distributed over the support $[\underline{x}, \overline{x}]$ according to the density function $f(x)$.[16] 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. The weight given to the appropriate action, which can be interpreted as the user's level of attention, is denoted m. Specifically, in our baseline model, we assume that when the user uses the AI, m is a decreasing function of the AI's autonomy (π), and that its shape is exogenously given (this assumption is relaxed in Section 6.2) and known by both the policymaker and the AI manufacturer. The assumption that the user attention decreases with AI autonomy can be interpreted as passive fatigue (Desmond and Hancock, 2001; Saxby et al., 2013). To illustrate, suppose that the user knows that the autonomy of an AI increases after the implementation of a better operating system (higher π). As a result, it will be more difficult (and less useful, see Section 6.2) for her to maintain attention.

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

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

Since $m(\pi) \geq 0$ 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 :

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

As inattention increases, the user's perception of the correct action tends to become more biased toward the default x^d.

If the user has to intervene, she chooses an action a and the harm $h(a, x)$ is realized. This harm is an increasing function of the distance between the user's chosen action a and the appropriate action

[16] We assume that f is a non-degenerate distribution. Specifically, f is continuous and strictly positive on support $[\underline{x}, \overline{x}]$.

x.[17] Specifically, when her intervention is required, the user chooses an action a that minimizes her *perceived* expected liability cost:[18]

$$\frac{\partial(\gamma h(a, x^s(m)))}{\partial a} = 0 \Leftrightarrow a = mx + (1-m)x^d \equiv a^s(x; m) \tag{2}$$

Human intervention is costly because (i) it requires a costly cognitive and/or physical effort, and (ii) the action chosen by the user is less appropriate than the one chosen by the AI. Regarding (i), the cost of cognitive and/or physical effort to the user in the case of an intervention is $k > 0$.[19] Indeed, as emphasized by Zerilli et al. (2019) and Walker et al. (2015), it is often very difficult to switch from low cognitive effort to high cognitive effort. Regarding (ii), the user's action is less appropriate than the AI's action due to behavioral inattention. To capture in a simple way the idea that the AI makes more accurate decisions, we assume that when the AI is able to act autonomously, it always chooses the appropriate action (there is no harm). Thus, the expected harm resulting from inattention is $H(m(0))$ if the AI is not used, and $(1-\pi)H(m(\pi))$ otherwise, with:

$$H(m) = \int_{\underline{x}}^{\overline{x}} f(x)h(a^s(x; m), x)dx \tag{3}$$

We make the following assumption:

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

Part (i) of Assumption 2 states that if the user is paying perfect attention, she will always choose the appropriate action and no harm will result. Part (ii) means that as the level of attention increases, the expected harm decreases. In the following, unless otherwise noted, we assume that the user has some degree of behavioral inattention, with $m(\pi) < 1 \ \forall \pi$, which implies that $H(m(\pi)) > 0 \ \forall \pi$.

Awareness of one's own inattention. We assume that the human user can be one of two types:

[17] For example, the damage function can be quadratic, with $h(a, x) = \frac{1}{2}(a-x)^2$. Other specifications can be used without changing our results, as long as Assumption 2 holds.

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

[19] This cost is independent of the parameter m and the type of the user (naive or sophisticated). If the user decides not to buy the AI, she always bears the cost k. Conversely, if she buys the AI, the cost k is discounted by $(1-\pi)$.

"naive" (with probability β) or "sophisticated" (with probability $1-\beta$). The human user type is unobservable to both the policymaker (when choosing the liability rule) and the AI manufacturer (when choosing the AI's autonomy and price). However, when choosing their actions, we assume that the policymaker and the AI manufacturer know the distribution of the user types (*i.e.*, β). A naive user is overconfident in her ability to avoid distractions and/or underestimates the risks resulting from her inattention. Formally, we assume that the naive user is unaware that she is subject to behavioral inattention and believes that $m(\pi) = 1 \; \forall \pi$. Consequently, her expectation of harm in the event of an intervention is always $H(1) = 0$. A sophisticated user, on the other hand, is well aware of her limited attentional capacity ($m(\pi) < 1$) and its potentially harmful consequences ($H(m(\pi)) > 0$).

The sequence of events. The timing of the game is as follows:

Stage 0. Nature chooses the type of the user (naive or sophisticated).

Stage 1. The policymaker chooses a liability sharing rule $\gamma \in [0, 1]$.

Stage 2. The AI manufacturer chooses whether to produce the AI, and if so, the level of AI autonomy $\pi \in [0, 1)$ and the price p at which the AI is sold.

Stage 3. The human user (who perfectly observes the AI's level of autonomy) makes two choices:

1. she decides whether to buy the AI and,
2. if the AI is not being used or requests the human user to step in, she intervenes by choosing an action a.

5 Analysis

We provide the first-best optimum in Section 5.1. We then solve the model by backward induction, starting with the human user's decision whether to buy the AI in Section 5.2. Next, we solve for the pricing decision and the AI autonomy chosen by the AI manufacturer in Section 5.3. Finally, we derive the optimal liability sharing rule in Section 5.4.

5.1 The first-best optimum

In this section, we derive the first-best optimum *conditional on user inattention*. This first-best minimizes the expected social cost, defined as the sum of the expected cost of human intervention and the cost of AI autonomy. We assume that using the AI is socially beneficial because it avoids a costly human intervention. Thus, at the first-best, all users should have access to the algorithm.

Recall that human intervention is costly because (i) it requires a cognitive and/or physical effort with cost k, and (ii) the action chosen by the user is less appropriate than the one chosen by the AI due to inattention. The expected harm resulting from inattention is $H(m(0))$ if the AI is not used, and $(1-\pi)H(m(\pi))$ otherwise.

When the AI is used, the expected harm $(1-\pi)H(m(\pi))$ varies ambiguously with AI autonomy. Specifically, it decreases with AI autonomy if:

$$(1-\pi)H'(m(\pi))m'(\pi) - H(m(\pi)) < 0 \tag{4}$$

A higher π has two opposite effects on the expected harm, respectively captured by the two terms on the left side of (4). First, by decreasing the user's level of attention, a higher π increases the harm in the event of an intervention, thus increasing the expected harm. Second, by reducing the probability of user intervention, a higher π reduces the probability that the user will cause harm: the expected harm decreases. The second effect dominates when condition (4) holds, and vice versa.[20]

Intuitively, the first effect is likely to dominate in contexts where it is very difficult for users to maintain attention as their active involvement decreases (automation creates a lot of passive fatigue), the expected harm from an intervention is low, and the information structure is complex.[21]

[20] Note that when $\pi = 0$, the expected harm $(1-\pi)H(m(\pi))$ unambiguously decreases with the AI autonomy.

[21] 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 is used.

For now, we assume that AI autonomy always reduces expected harm (this assumption will be relaxed in Section 6.1):

Assumption 3. *The expected harm $(1-\pi)H(m(\pi))$ always decreases with the AI autonomy (i.e., condition (4) is satisfied for all π):*

$$\frac{\mathrm{d}(1-\pi)H(m(\pi))}{\mathrm{d}\pi} < 0 \; \forall \pi \tag{5}$$

Recall that the cost of developing an AI is $c(\pi)$. If everyone uses the AI, then the expected social cost is:

$$SC(\pi) = (1-\pi)(k + H(m(\pi))) + c(\pi) \tag{6}$$

Assuming an interior solution, the first-best level of autonomy is characterized by the following first-order condition (FOC):[22]

$$\frac{\partial SC}{\partial \pi}(\pi) = 0 \Leftrightarrow k - \frac{\mathrm{d}\left[(1-\pi)H(m(\pi))\right]}{\mathrm{d}\pi} = c'(\pi) \tag{10}$$

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 (10).*

[22] The second order condition (SOC) is:

$$c''(\pi) > 2H'(m(\pi))m'(\pi) - (1-\pi)(H''(m(\pi))m'(\pi) + H'(m(\pi))m''(\pi)) \tag{7}$$

With:

$$H'(m(\pi)) = \int_0^1 f(x)\frac{\partial h}{\partial a}(a^s(x;m(\pi)),x)\frac{\partial a^s}{\partial m}(x;m(\pi))dx \tag{8}$$

And:

$$H''(m(\pi)) = \int_0^1 f(x)\left(\frac{\partial h}{\partial a}(a^s(x;m(\pi)),x)\frac{\partial^2 a^s}{\partial m^2}(x;m(\pi)) + \frac{\partial^2 h}{\partial a^2}(a^s(x;m(\pi)),x)\left(\frac{\partial a^s}{\partial m}(x;m(\pi))\right)^2\right)dx \tag{9}$$

Furthermore, it is easy to show that condition (7) also satisfies the AI manufacturer's SOCs associated with the FOCs (19) and (21). In the following, we assume that (7) is satisfied.

Next, we solve the model by backward induction, starting with the human user's decision whether to buy the AI.

5.2 Stage 3: the human user's decision to purchase the AI.

The naive user. Recall that the naive user's lack of awareness of behavioral inattention implies that $m(\pi) = 1 \ \forall \pi$. Consequently, her expectation of harm in the event of an intervention is $H(1) = 0$, and the only reason why a naive user might want to buy the AI is to reduce her expected cost of intervention, for an expected savings of πk. She will therefore choose to buy the AI if the price p is such that:

$$p \leq \pi k \equiv \underline{p}(\pi) \tag{11}$$

Due to the naive user's misperception of her level of attention, her willingness to pay $\underline{p}(\pi)$ ignores the fact that the AI outperforms her: the expected harm of an intervention does not affect her decision to buy the AI.

The sophisticated user. The sophisticated user, on the other hand, is well aware of her limited level of attention. When using an AI, she expects she might have to intervene with probability $(1 - \pi)$, and will cause an expected harm $H(m(\pi))$. She bears a fraction γ of that expected harm. If she does not use the 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). Finally, the sophisticated user buys the AI when:

$$(1 - \pi)[k + \gamma H(m(\pi))] + p \leq k + H(m(0)) \tag{12}$$

Which is equivalent to:

$$p \leq \pi \left[k + \gamma H(m(\pi))\right] + (1 - \gamma)H(m(\pi)) - [H(m(\pi)) - H(m(0))] \equiv \overline{p}(\pi, \gamma) \tag{13}$$

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 $\gamma < 1$, 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, according to Assumptions 1 and 2, the sophisticated user expects that when the AI is used, her attention will be lower, and therefore the expected damage in case of an intervention will be higher.

These two opposing effects make it seemingly possible that, if the difference $H(m(\pi)) - H(m(0))$ is large enough, the willingness to pay of sophisticated users will be lower than that of naive users ($\overline{p}(\pi, \gamma) \leq \underline{p}(\pi)$).[23] However, Assumption 3 guarantees that this is not the case, as specified in the following lemma.

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

Proof. If $\pi = 0$, the willingness to pay of the sophisticated user is $\overline{p}(0, \gamma) = (1 - \gamma)H(m(0)) \geq 0$ 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:

$$\overline{p}(0, \gamma) \geq \underline{p}(0) \tag{14}$$

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

$$\overline{p}(\pi, \gamma) = \underline{p}(\pi) - \gamma(1 - \pi)H(m(\pi)) + H(m(0)) \tag{15}$$

From which:

$$\frac{\partial \overline{p}}{\partial \pi}(\pi, \gamma) = \underline{p}'(\pi) - \gamma \frac{\mathrm{d}\left[(1 - \pi)H(m(\pi))\right]}{\mathrm{d}\pi} \tag{16}$$

[23] 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.

With $\underline{p}'(\pi) = k > 0$. It follows from Assumption 3 and (16) that:

$$\frac{\partial \overline{p}}{\partial \pi}(\pi, \gamma) \geq \underline{p}'(\pi) \tag{17}$$

From (14) and (17), and since $\underline{p}$ and $\overline{p}$ are continuous with respect to π, we can deduce that $\overline{p}(\pi, \gamma) \geq \underline{p}(\pi)\ \forall(\pi, \gamma)$. □

5.3 Stage 2: The AI manufacturer's pricing decision and investment in AI autonomy

Recall that the AI manufacturer is assumed to be a monopolist.[24] 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 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 $\overline{p}(\pi, \gamma)$ 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). His expected profit is:

$$\Pi_{\underline{p}}(\pi, \gamma) = \pi k - (1 - \pi)(1 - \gamma)H(m(\pi)) - c(\pi) \tag{18}$$

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

$$\frac{\partial \Pi_{\underline{p}}}{\partial \pi}(\pi, \gamma) = 0 \Leftrightarrow k - (1 - \gamma)\frac{\mathrm{d}\left[(1 - \pi)H(m(\pi))\right]}{\mathrm{d}\pi} = c'(\pi) \tag{19}$$

[24] 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.

We denote by $\pi^*_{\underline{p}}(\gamma)$ 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 (19)). 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 (19)).

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

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

Now let us assume that the AI manufacturer sets the price $\overline{p}(\pi, \gamma)$ (the AI is sold only to sophisticated users). The AI manufacturer's expected profit, after simplification, is:

$$\Pi_{\overline{p}}(\pi) = (1-\beta)\left[\pi(k + H(m(\pi))) - (H(m(\pi)) - H(m(0)))\right] - c(\pi) \tag{20}$$

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

$$\frac{\partial \Pi_{\overline{p}}}{\partial \pi}(\pi) = 0 \Leftrightarrow (1-\beta)\left[k - \frac{\mathrm{d}\left[(1-\pi)H(m(\pi))\right]}{\mathrm{d}\pi}\right] = c'(\pi) \tag{21}$$

We denote by $\pi^*_{\overline{p}}$ the degree of autonomy implicitly defined by this FOC. The liability sharing rule does not affect the choice of π in this case. Indeed, the manufacturer's share of the damage decreases with π, either indirectly through the price (for a fraction γ of the damage), or directly through the liability sharing rule (for a fraction $1-\gamma$ of the damage). Moreover, the manufacturer's marginal revenue is now 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

$\overline{p}(\pi,\gamma)$? It depends mainly on the relative probability of being confronted with a naive user versus a sophisticated one. To understand why, let us first assume that $\beta = 0$ (*i.e.*, all users are sophisticated). In this case, $\Pi_{\overline{p}}(\pi) > \Pi_{\underline{p}}(\pi,\gamma)$ for all levels of AI autonomy, and the AI manufacturer is better off choosing the "high" price $\overline{p}(\pi^*_{\overline{p}},\gamma)$. Now let us assume that $\beta = 1$ (*i.e.*, all users are naive). In this case, we have $\Pi_{\overline{p}}(\pi) \leq 0$, and the manufacturer may prefer to sell at the "low" price $\underline{p}(\pi^*_{\underline{p}}(\gamma))$.[25]

5.4 Stage 1: The policymaker's choice of a liability sharing rule

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

5.4.1 Optimal sharing rule for given prices

First, we note that in the absence of behavioral inattention ($m(\pi) = 1$), users will always choose the action that is objectively the most appropriate ($a = x$), regardless of the level of AI autonomy. 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.[26]

Now suppose the AI manufacturer chooses a "low" price $\underline{p}(\pi^*_{\underline{p}}(\gamma))$ so that all users buy the AI. A comparison using the FOCs (10) and (19) shows that $\pi^*_{\underline{p}}(\gamma)$ is inferior to the first-best, unless the AI manufacturer is fully liable for the damage ($\gamma = 0$).[27] Otherwise (if $\gamma > 0$), the expected liability cost faced by the manufacturer does not allow him to fully internalize the marginal social benefit, in terms of reduced expected harm, of increasing the autonomy of the AI. As a result, the manufacturer underinvests in autonomy.

Finally, suppose the AI manufacturer chooses a "high" price $\overline{p}(\pi^*_{\overline{p}},\gamma)$ so that only sophisticated

[25] However, as explained in Section 5.4.2, the AI manufacturer may not be able to obtain a positive expected profit by choosing the price $\underline{p}(\pi^*_{\underline{p}}(\gamma))$ in this circumstance.

[26] 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 $\overline{p}(\pi,\gamma)$ converge: $\underline{p}(\pi) = \overline{p}(\pi,\gamma) = \pi k$.

[27] Note that when $\gamma = 0$, the user does not necessarily choose the most appropriate action according to her 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 she chooses $a = a^s(x;m)$.

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, the manufacturer's choice of AI autonomy is socially optimal. Indeed, if the price chosen by the AI manufacturer is such that only sophisticated users buy the AI, the expected social cost is:

$$SC_+(\pi) = (1-\beta)(1-\pi)(k + H(m(\pi))) + \beta(k + H(m(0))) + c(\pi) \tag{22}$$

The level of AI autonomy that minimizes $SC_+(\pi)$ is characterized by a FOC equivalent to (21), which characterizes the level of autonomy chosen by the manufacturer ($\pi^*_{\overline{p}}$). Note also that the level of AI autonomy and the expected social cost are independent of the liability sharing rule (γ).

The following proposition summarizes the previous findings.

Proposition 2. *If the manufacturer chooses the "low" price $\underline{p}(\pi^*_{\underline{p}}(\gamma))$, the AI autonomy is socially optimal only if $\gamma = 0$. If the manufacturer chooses the "high" price $\overline{p}(\pi^*_{\overline{p}}, \gamma)$, 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.

5.4.2 Optimal sharing rule with endogenous pricing

So far, we have focused on the optimal liability sharing rule separately for prices $\underline{p}(\pi^*_{\underline{p}}(\gamma))$ and $\overline{p}(\pi^*_{\overline{p}}, \gamma)$. 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 (γ) on the manufacturer's pricing decision. Using the envelope theorem, we find:

$$\frac{\mathrm{d}\Pi_{\underline{p}}(\pi^*_{\underline{p}}(\gamma), \gamma)}{\mathrm{d}\gamma} = (1 - \pi^*_{\underline{p}}(\gamma))H(m(\pi^*_{\underline{p}}(\gamma))) > \frac{\mathrm{d}\Pi_{\overline{p}}(\pi^*_{\overline{p}})}{\mathrm{d}\gamma} = 0 \tag{23}$$

Thus, as the share of liability borne by the user (γ) increases, it becomes relatively more profitable for the manufacturer to charge a "low" price $\underline{p}(\pi^*_{\underline{p}}(\gamma))$. This is because the naive user's willingness to pay does not decrease with her share of liability, while it reduces the manufacturer's direct liability cost.

Proposition 3. *Increasing the user's share of liability (γ) may induce the manufacturer to lower its price.*[28]

Thus, increasing the user's share of liability (γ) may be socially beneficial in that naive users will then use the AI, reducing the total expected cost through both the expected cost of user intervention and the expected harm.

Another advantage of choosing a high enough user liability is that if the liability falls mainly on the AI manufacturer, it may discourage AI development altogether. Indeed, the AI manufacturer may not be able to make a positive expected profit. To understand why, consider the AI manufacturer's expected profit with prices $\overline{p}(\pi^*_{\overline{p}}, \gamma)$ and $\underline{p}(\pi^*_{\underline{p}}(\gamma))$, respectively. First, if the AI manufacturer chooses the price $\overline{p}(\pi^*_{\overline{p}}, \gamma)$, the presence of a fixed cost $c(0) > 0$ implies that his expected profit is strictly negative if the proportion of naive users (β) is high enough.[29] Second, if the AI manufacturer chooses the price $\underline{p}(\pi^*_{\underline{p}}(\gamma))$, Lemma 2 implies that the AI manufacturer's expected profit decreases with his share of liability, up to the point where it eventually becomes negative. Thus, if the proportion of naive users is high (β is high) and the AI manufacturer is liable for a significant fraction of the expected harm (γ is low), he may choose not to develop the AI. This decision is socially costly (recall that, by assumption, developing and distributing the AI to all users is socially beneficial).

In summary, because some users may be prone to some degree of behavioral inattention without

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

[29] The existence of a fixed cost is a necessary condition for this to be true. Suppose there is no fixed cost ($c(0) = 0$). Under Assumption 3, it is then possible to show that the AI manufacturer can always obtain a positive expected profit by choosing a high price (except in the extreme case where $\beta = 0$, in which case the AI manufacturer's expected profit is 0). This is because, if $c(0) = 0$, then $\Pi_{\overline{p}}(0) = 0$ from (20), and $\Pi'_{\overline{p}}(0) > 0$ from (21) and Assumption 3. Thus, without fixed costs, it is always advantageous for the AI manufacturer to develop the AI in order to sell it at least to sophisticated users: the AI is always developed when $c(0) = 0$.

being aware of it, the policymaker may face a trade-off when setting the liability sharing rule.[30] 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 helps to avoid situations in which the AI manufacturer does not develop or sell the AI, and if the AI is actually developed and sold, it incentivizes the AI manufacturer to lower the price. Thus, a larger fraction of users will benefit from the AI.[31]

6 Extensions

In Section 6.1, we discuss how our results are affected when we relax Assumption 3. In Section 6.2, we endogenize the choice of attention by assuming that users can increase their attention through costly cognitive effort.

6.1 Expected harm may increase with AI autonomy

We relax Assumption 3 and instead assume that the expected harm $(1-\pi)H(m(\pi))$ first decreases with π (the effect on the probability of user intervention dominates), and then increases above a certain level of π (the effect on user inattention dominates). Furthermore, the expected damage is assumed to be convex with respect to π. Following a line of reasoning similar to that of Dawid et al. (2024), our results will vary significantly depending on the shape of the cost function $c(\pi)$ and the magnitude of the cost k.

[30] 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.

[31] Note that our results are based on the assumption that $k > 0$. However, as one reviewer of this paper pointed out, some individuals may choose not to use the AI because they want to retain control, in which case it could be assumed that $k < 0$ (the cost of an intervention to a user is less than the benefit of retaining control). Let us consider how this assumption would affect our results. The willingness to pay for the AI of naive users would be strictly negative. As a result, the manufacturer would only meet the demand of sophisticated users (at most): the unawareness of naive users of their inattention could lead to an insufficient diffusion of the AI. The sophisticated users will be willing to buy the AI only if it sufficiently reduces their expected liability costs (a necessary condition for the development of the AI). If this is the case, and the AI manufacturer can earn positive expected profits by developing the AI, then under Assumption 3 our results are similar to those obtained when the AI manufacturer sets the price to $\overline{p}(\pi, \gamma)$: regardless of the liability sharing rule, the level of AI autonomy is socially optimal given that only sophisticated users use the AI.

Suppose that the marginal cost of AI autonomy increases rapidly while the cost k is low, so that the socially optimal level of AI autonomy is relatively low and, as a consequence, the expected harm decreases with the level of AI autonomy (the reduced probability of user intervention effect dominates), as in Assumption 3. In this case, our results are unchanged: the willingness to pay of sophisticated users is higher than that of naive users, and the level of AI autonomy chosen by the AI manufacturer may be too low. There is still a trade-off between incentivizing the AI manufacturer to increase its investment in AI autonomy on the one hand, and improving AI diffusion on the other hand.

Now suppose that the marginal cost of AI autonomy increases slowly while the cost k is high, so that at the socially optimal level of autonomy, the expected harm increases with the level of AI autonomy (the inattention effect dominates).[32] To make it easier to derive some intuitions, we consider only two polar cases: (i) the AI manufacturer faces strict liability ($\gamma = 0$), and (ii) the user faces strict liability ($\gamma = 1$).

Let us first consider case (i) ($\gamma = 0$). Regardless of the autonomy of the AI, the willingness to pay of the sophisticated user is higher than that of the naive user. This is because, unlike the naive user, the sophisticated user takes into account the fact that if she uses the AI, the damage will be borne entirely by the manufacturer. The price depends on the probability that the user is sophisticated. If the probability is low, the manufacturer chooses a price such that the AI will be purchased by all users. Conversely, if the probability is high, the manufacturer sets the price so that only sophisticated users will buy the AI. The diffusion of the AI may thus be insufficient. However, for a given price (and thus for a given AI diffusion), strict liability of the manufacturer implies that he fully internalizes the harm and chooses the socially optimal level of AI autonomy.

Now consider case (ii) ($\gamma = 1$). If the inattention effect is large enough, there is a threshold of AI

[32] Note that this is a rather extreme case, since at the level of autonomy chosen by the AI manufacturer, reducing that autonomy at the margin would reduce the expected harm. Such a situation is likely to raise some social acceptance issues. To illustrate, this could correspond to a situation where the autonomy level of autonomous vehicles is chosen to allow drivers to use their driving time for other tasks (high autonomy reduces the expected cost of interventions), even though requiring a more active role from the driver (by reducing the vehicle's autonomy level) would reduce the probability and/or severity of accidents.

autonomy above which the sophisticated user's willingness to pay becomes lower than that of the naive user.

First, when the autonomy level chosen by the AI manufacturer is below that threshold, our results remain largely unchanged. The main difference is that, when the AI is sold at a low price (now equal to the naive user's willingness to pay), the AI autonomy chosen by the manufacturer tends to be too high, since the marginal effect of autonomy on expected harm is now positive. Increasing the manufacturer's liability pushes his choice of AI autonomy towards the socially optimal level, but has a negative effect on his expected profit when the price is low, encouraging him to distribute the AI less widely. In other words, we find that there is still a trade-off between providing the right incentives to the manufacturer (but now to reduce the AI autonomy), and promoting AI diffusion.

Second, if the level of autonomy chosen by the AI manufacturer is so high that the sophisticated user's willingness to pay is *lower* than that of the naive user (the inattention effect is very strong), the AI manufacturer sets a price equal to the naive user's willingness to pay. Indeed, the AI manufacturer can then both charge a high price and sell the AI to all types of users. Again, increasing the manufacturer's share of liability pushes his choice of AI autonomy towards the socially optimal level. However, as the manufacturer's liability increases (and the user's liability decreases), the sophisticated user's willingness to pay increases and may exceed that of the naive user. In this case, we return to the diffusion problem: the AI may only be sold to the sophisticated user.

6.2 The cost of attention

In this extension, we assume that the user's level of attention m is a choice variable that can be increased through a costly effort (the choice is made before the user may have to intervene).[33] 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.

[33]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).

Regarding the timing of the game, we add an additional stage (stage 4) in which the human user, after deciding whether to buy the AI in stage 3, chooses her level of attention.

6.2.1 Stage 4: The human user's choice of attention

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:

$$(1 - \pi)\gamma H(m) + c_o(m) + p \tag{24}$$

The FOC for the level of attention is:

$$-(1 - \pi)\gamma H'(m) = c_o'(m) \tag{25}$$

We denote by $\underline{m}(\pi, \gamma)$ the level of attention of the user characterized by this FOC. Increasing her level of attention is (cognitively) costly (right-hand side of (25)), but it reduces the expected harm, and thus her expected liability cost (left-hand side of (25)). 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 (γ):[34]

$$\frac{\partial \underline{m}}{\partial \pi}(\pi, \gamma) \leq 0, \text{ and } \frac{\partial \underline{m}}{\partial \gamma}(\pi, \gamma) > 0 \text{ if } \pi < 1 \tag{26}$$

Moreover:

$$\frac{\partial^2 \underline{m}}{\partial \pi \partial \gamma}(\pi, \gamma) < 0 \tag{27}$$

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

[34] As in the baseline model, 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 $\gamma < 1$.

AI.[35]

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

$$H(m) + c_o(m) \tag{28}$$

The FOC for the level of attention is:

$$H'(m) = c_o'(m) \tag{29}$$

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 $\overline{m} \geq \underline{m}(\pi, \gamma)$: the level of attention chosen by the user is higher when she is not using the AI. These results are generally consistent with the assumptions made in the baseline model, except that the user's level of attention when using an AI now increases with the user's liability.

6.2.2 Stage 3: The human user's purchase decision

The user will acquire the AI if:

$$p \leq [c_o(\overline{m}) - c_o(\underline{m}(\pi, \gamma))] + [H(\overline{m}) - (1-\pi)\gamma H(\underline{m}(\pi, \gamma))] \equiv p^*(\pi, \gamma) \tag{30}$$

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

6.2.3 Stage 2: The manufacturer's choice of AI autonomy

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

$$\Pi(\pi, \gamma) = [c_o(\overline{m}) - c_o(\underline{m}(\pi, \gamma))] + [H(\overline{m}) - (1-\pi)H(\underline{m}(\pi, \gamma))] - c(\pi) \tag{31}$$

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

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. Note that the expected harm is fully internalized by the manufacturer (through the combination of price and manufacturer liability).

The FOC for the AI autonomy is:

$$\frac{\partial \Pi}{\partial \pi}(\pi, \gamma) = H(\underline{m}(\pi, \gamma)) - c'(\pi) - \frac{\partial \underline{m}}{\partial \pi}(\pi, \gamma)[(1-\pi)H'(\underline{m}(\pi, \gamma)) + c_o'(\underline{m}(\pi, \gamma))] = 0 \tag{32}$$

Substituting the user's FOC (25) into (32), we get:

$$\frac{\partial \Pi}{\partial \pi}(\pi, \gamma) = H(\underline{m}(\pi, \gamma)) - c'(\pi) - \frac{\partial \underline{m}}{\partial \pi}(\pi, \gamma)(1-\pi)(1-\gamma)H'(\underline{m}(\pi, \gamma)) = 0 \tag{33}$$

We denote by $\pi^*(\gamma)$ the AI autonomy chosen by the manufacturer characterized by (33). 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 $\Pi(\pi^*(\gamma), \gamma) > 0 \ \forall \gamma$ (*i.e.*, the manufacturer is always willing to develop and market the AI), which implies that $p^*(\pi^*(\gamma), \gamma) > 0 \ \forall \gamma$.

6.2.4 Stage 1: The socially optimal sharing of liability

The expected social cost is:

$$SC_*(\gamma) = (1 - \pi^*(\gamma))H(\underline{m}(\pi^*(\gamma), \gamma) + c_o(\underline{m}(\pi^*(\gamma), \gamma) + c(\pi^*(\gamma)) \tag{34}$$

After simplification and using (33) and (25), the FOC for the optimal liability sharing rule can be written:

$$\frac{\partial SC_*}{\partial \gamma}(\gamma) = \frac{\partial \underline{m}}{\gamma}(\pi^*(\gamma), \gamma)(1 - \pi^*(\gamma))(1-\gamma)H'(\underline{m}(\pi^*(\gamma), \gamma)) = 0 \tag{35}$$

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

Proposition 4. *Assume that the user can increase her level of attention m at a cost $c_o(m)$ and*

that all users are sophisticated ($\beta = 0$). No liability of the manufacturer ($\gamma = 1$) is socially optimal.

The intuition is as follows. Compared to the baseline model, since the user is always of the sophisticated type, increasing the manufacturer's share of liability is not socially beneficial, 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 extension). 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 that includes both sophisticated and naive users, as well as the possibility of costly attentional effort on the part of users, we can expect a more nuanced trade-off between (i) increasing the manufacturer's share of liability to improve the degree of AI autonomy, and (ii) increasing the user's share of liability to improve both AI diffusion and user attentional effort.

7 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 also assumed that the performance of the AI is better than that of a human user, and that users may be subject to behavioral inattention (Gabaix, 2019). Inattention, which is expected to increase with AI autonomy, can lead to poor decisions and 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 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.

We believe that our work has several important policy implications. We show that inattention, coupled with some users' lack of awareness of their attentional limitations, may limit the diffusion of AI technologies. Several legal scholars have already suggested that manufacturer liability may delay the introduction of AI technologies (see, *e.g.*, Parchomovsky and Stein, 2008; Schellekens, 2015; Dawid and Muehlheusser, 2022).[36] We extend this argument by showing that inattention may provide an additional incentive not to bring the technology into the market, even if it could perform reasonably well (in terms of improving expected social welfare). Moreover, our results show that liability affects not only the market introduction of the innovation, but also the extent to which the innovation is used. When allocating liability, policymakers should consider the consequences of user inattention: increasing manufacturer liability may limit the diffusion of AI, thereby limiting its socially beneficial effects. In addition to encouraging better adoption of AI, tilting the balance toward less manufacturer liability (and more user liability) has the added benefit of increasing the incentive for users to be less distracted. Unfortunately, reduced manufacturer liability comes at a cost, as the manufacturer's investment in AI autonomy tends to be suboptimal. Thus, when choosing the liability sharing rule, policymakers face a trade-off between (i) supporting AI diffusion and increasing the attention level of the users, and (ii) incentivizing the manufacturer to invest more in the AI autonomy.[37] Although we do not provide a dynamic framework, one possible way for policymakers to address this trade-off is to focus on AI diffusion as a first step (by reducing manufacturer liability) and then, in a second step (once the AI has been adopted by a significant

[36] Some innovations may significantly reduce the risk and/or severity of accidents, while others may provide important user benefits despite a constant or even higher risk. Liability will have a different impact on the introduction of each type of innovation (see Galasso and Luo, 2017).

[37] Note that similar policy recommendations have previously been proposed in the literature focusing on liability and the timing of innovation (Dawid and Muehlheusser, 2022), although we are the first to focus on behavioral inattention (to the best of our knowledge).

fraction of potential users), to increase manufacturer liability in order to provide better investment incentives.[38]

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).[39] Finally, to the extent that the manufacturer is able to sell the use of its AI to sophisticated users at a higher price, it may be in his interest to 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*

[38] Another benefit of focusing on AI diffusion as a first step is that users can learn to better assess how inattention affects the likelihood and severity of accidents.

[39] 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).

International Conference, SAFECOMP 2009, Hamburg, Germany, September 15-18, 2009. Proceedings 28, pages 18–31. Springer.

Andrews, S., Ellis, D. A., Shaw, H., and Piwek, L. (2015). Beyond self-report: Tools to compare estimated and real-world smartphone use. *PLoS ONE*, 10(10):e0139004.

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.

Cunningham, M. L. and Regan, M. A. (2018). Driver distraction and inattention in the realm of automated driving. *IET Intelligent Transport Systems*, 12(6):407–413.

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., Di, X., Kort, P. M., and Muehlheusser, G. (2024). Autonomous vehicles policy and safety investment: an equilibrium analysis with endogenous demand. *Transportation Research Part B: Methodological*, 182:102908.

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*, 134:104288.

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*, 68:106022.

Desmond, P. A. and Hancock, P. A. (2001). Active and passive fatigue states. In *P. A. Hancock & P. A. Desmond (Eds.), Stress, workload, and fatigue*, pages 455–465. Lawrence Erlbaum Associates Publishers.

Di, X., Chen, X., and Talley, E. (2020). Liability design for autonomous vehicles and human-driven

vehicles: A hierarchical game-theoretic approach. *Transportation research part C: emerging technologies*, 118:102710.

Feess, E. and Muehlheusser, G. (2024). Autonomous vehicles: Moral dilemmas and adoption incentives. *Transportation Research Part B: Methodological*, 181:102894.

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.

Galasso, A. and Luo, H. (2017). Tort reform and innovation. *The Journal of Law and Economics*, 60(3):385–412.

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

Klauer, S. G., Dingus, T. A., Neale, V. L., Sudweeks, J. D., and Ramsey, D. J. (2006). The impact of driver inattention on near-crash/crash risk: an analysis using the 100-car naturalistic study data. Technical report, National Highway Traffic Safety Administration.

Knowles, D. and Tay, R. S. (2002). Driver inattention: More risky than the fatal four? In *2002 Road Safety Research, Policing and Education Conference, Adelaide*, pages 87–91. Transport SA, Australia.

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.

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

Parchomovsky, G. and Stein, A. (2008). Torts and innovation. *Michigan Law Review*, 107(2):285–315.

Parry, D. A., Davidson, B. I., R., S. C. J., Fisher, J. T., Mieczkowski, H., and S., Q. D. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nat Hum Behav*, 5(11):1535–1547.

Saxby, D. J., Matthews, G., Warm, J. S., Hitchcock, E. M., and Neubauer, C. (2013). Active and passive fatigue in simulated driving: discriminating styles of workload regulation and their safety impacts. *Journal of Experimental Psychology: Applied*, 19(4):287–300.

Schellekens, M. (2015). Self-driving cars and the chilling effect of liability law. *Computer Law & Security Review*, 31(4):506–517.

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.

Walker, G., Stanton, N., and Salmon, P. (2015). *Human Factors in Automotive Engineering and Technology*. Human Factors in Transport Series. Ashgate Publishing Ltd.

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