Mitigating Generative AI Hallucinations*

Alessandro De Chiara†, Ester Manna‡, and Shubhranshu Singh§

March 23, 2026

Abstract

We theoretically investigate whether AI developers or AI operators should be liable for the harm the AI systems may cause when they hallucinate. We find that the optimal liability framework may vary over time, with the evolution of the AI technology, and that making the AI operators liable can be desirable only if it induces monitoring of the AI systems. We also highlight non-trivial relationships between welfare and reputational concerns, human supervision ability, and the accuracy of the technology. Our results have implications for regulatory design and business strategies.

Keywords: AI hallucinations; AI liability; AI supervision

JEL classifications: K2; L51.

*We would like to thank Juan-José Ganuza, Eric Langlais, Andrea Mantovani, and Kathryn E. Spier, as well as the audiences at the Law, Institutions and Economics seminar at the Université Paris Nanterre (France) and at the IV UB Microeconomics Workshop (Spain) for valuable comments and helpful discussions. This work has been financially supported by the Spanish Ministry of Science, Innovation and Universities, Spain through grants PID2020-114040RB-I00 and PID2021-128237OB-I00, and by the Catalan Research Agency (grant number 2021-SGR00678). The usual disclaimer applies.

†Department of Economics, Universitat de Barcelona and Barcelona Economic Analysis Team (BEAT), Avinguda Diagonal 696, 08034, Barcelona, Spain. E-mail: aledechiara@ub.edu.

‡Professora Lectora Serra Húnter, Department of Economics, Universitat de Barcelona and Barcelona Economic Analysis Team (BEAT), Avinguda Diagonal 696, 08034, Barcelona, Spain. E-mail: estermanna@ub.edu.

§Carey Business School, Johns Hopkins University, 100 International Drive, Baltimore, MD 21202. E-mail: shubhranshu.singh@jhu.edu.

1 Introduction

ChatGPT's 2022 launch stunned the world and laid bare the transformative potential of *generative AI*, namely, AI that generates new content, such as text, images, lines of computer code, or music. This development comes with significant perils because AI models are plagued by *hallucinations*: being designed to give answers that are statistically likely rather than factually correct, AI systems may generate confident and plausible but false outputs. An expert's decision to rely on a generative AI system exposes clients to risks of harm due to AI hallucinations, thus creating urgency around the question of how to effectively mitigate generative AI hallucinations.[1] In this paper, we study an AI firm's incentive to invest in reducing the likelihood of AI hallucinations and an expert's incentive to verify the exactness of the AI signal (so that hallucinations can be identified and ignored) under different policies for allocating liability for damages. The following examples highlight some of the important features that we capture in our model.

Physicians may use generative AI's assistance to diagnose patients. If the physician discovers that the AI-generated information is a hallucination (e.g., based on a fabricated research study), she ignores the AI generated information. However, if the physician fails to detect the AI hallucination, she may misdiagnose the patient and cause the patient harm. Legal professionals, such as lawyers and prosecutors, may seek generative AI's assistance in a wide range of tasks, for example, when deciding whether to settle a legal case or take it to court. A decision based on an undetected AI hallucination can potentially harm the defendant. Similarly, engineers and architects may decide to use generative AI for the creation and development of new software or for the design and prototyping of projects, such as buildings, bridges, or machines. These designs and projects may fail if the AI recommendation is based on a hallucination.[2]

[1] The now infamous case of a New York lawyer citing fabricated legal precedents from ChatGPT underscores the potential for real-world harm. See *"Here's What Happens When Your Lawyer Uses ChatGPT,"* The New York Times, May 27, 2023. Dahl et al. (2024) highlights the pervasiveness of legal hallucinations.

[2] For an incomplete list of possible applications and drawbacks, see "*Will generative AI transform business?*" published in the Financial Times on October 23, 2023. Concerning the prospects of using generative AI for medical advice and the design of building structures, see Haupt and Marks (2023) and Liao et al. (2024), respectively.

In the above examples, experts face two choices. First, they must decide whether to rely on generative AI, which is often subscription based, to carry out a task (e.g., diagnosing a patient or giving a legal advice). Second, if the expert decides to use AI, she must decide whether to verify if the AI system produced a hallucination. Learning that a hallucination occurred informs the expert that the AI-generated information is inexact and should be tossed away. However, if the expert fails to identify the hallucination, she risks making an incorrect decision and causing harm to clients. The question of whether the AI firm or the expert should bear the liability for damages when client harm occurs as a result of the expert relying on an AI hallucination to make decisions naturally arises.

Indeed, regulators and lawmakers around the world are racing to update regulations to catch up with the rapid advances of AI. The EU recently approved measures that include the AI Act and the revised Product Liability Directive, which applies strict liability principles to AI products.[3] However, the legal framework is not devoid of ambiguities, partly stemming from the withdrawal of the more general AI Liability directive, which would have introduced a presumption of fault for AI operators for a broad range of damages. In the US, regulatory intervention is mostly directed at minimizing cybersecurity threats, citizen deception, and ensuring workers' rights and fairness (e.g., ensuring that algorithms do not discriminate against specific groups).[4] While online platforms have been shielded from liability arising from third-party content thanks to Section 230 of the 1996 Communications Decency Act (CDA), whether artificial hallucinations would be protected according to the same provision is unclear. On the one hand, generative AI uses datasets and information collected online (hence, information provided by third parties). On the other hand, the hallucinations are created by the AI itself based on its own programming and training data. It follows that the people who created and trained the AI may be liable for any harm the artificial hallucination may cause.[5]

Despite good intentions, policy discussions without clear guidance and a proper un-

[3]Under strict liability, an injured party is not required to prove the producer's fault or negligence to obtain compensation. The revised Product Liability Directive broadens the definition of products, compared to Council Directive 85/374/EEC, to include software, AI systems, and similar digital goods.

[4]See President Biden's Executive Order 14110 on "*Safe, Secure, and Trustworthy Artificial Intelligence*", signed on October 30, 2023.

[5]See the report "*Section 230 Immunity and Generative Artificial Intelligence*" published by the Congressional Research Service on October 28, 2023.

derstanding of the effects of alternative liability regimes may lead to regulations that inadequately protect consumers and/or stifle innovation.[6] In an effort to address this gap, we develop a theoretical model that captures key elements of the interaction between generative AI and a human expert under different liability regimes. In the model, the expert decides whether to generate an AI signal (or recommendation) about whether to undertake an activity, which can potentially cause harm to the client. The expert has an inherent desire to avoid negative outcomes for the client, irrespective of how liability is allocated. We term this desire the expert's "*reputational concern.*" After receiving the AI signal, the expert can exert a costly effort to imperfectly verify the exactness of the signal. The likelihood that the AI signal is exact (or the AI reliability) depends on the baseline state of the technology as well as the AI firm's privately observed investment in accuracy.

In this environment, we show that assigning liability for harm to third parties to the DM ("*AI-Operator liability*") provides first-best incentives to select the activity and engage in AI supervision. Conversely, assigning liability to the AI developer ("*AI-Developer liability*") distorts such incentives. Intuitively, with AI-Operator liability, the DM may want to adopt the AI only if the recommendation can reveal information that may make him reevaluate the best course of action. With AI-Developer liability, there is an additional reason for why the DM may want to adopt the AI system: the AI not only provides useful information, but can also allow shifting liability if things go awry. Yet, the distortion is not always in the direction one would expect. We identify scenarios where AI-Developer liability paradoxically results in too little risk taking and reduced AI adoption relative to AI-Operator liability: the DM may be reluctant to use the AI, even though he can shift liability to the developer, anticipating that he will have little incentive to engage in supervision and that a negative outcome may damage his reputation (see Proposition 1).

When we include investment incentives, a trade-off arises between the two liability regimes because accountability is what motivates the AI developer to invest in accuracy.

[6]The measures proposed and adopted in the EU have led to some industry detractors to argue that they might hinder innovation, including the French president Emmanuel Macron (see "*EU's new AI Act risks hampering innovation, warns Emmanuel Macron*" published in the Financial Times on December 11, 2023).

This trade-off is resolved in favor of AI-Operator liability only when inducing AI supervision is critical. Moreover, we find that when the baseline accuracy of the technology is sufficiently advanced, AI supervision can be dispensed with, and making the AI developer liable can further improve the AI system reliability. This finding suggests that, as the state of the technology evolves, the optimal allocation of liability may change (see Proposition 2). The gist of these conclusions is not altered by the possibility of imposing the duty of supervising the AI onto the DM to escape liability ("*AI-Developer Conditional liability*"). This alternative regime, which can be demanding in terms of which information should be verifiable in court, further reduces the scope for adopting AI-Operator liability, which would be mostly confined to situations in which the DM's reputation does not play a relevant role in determining his choices (see Proposition 4). We also highlight some intriguing and non-trivial comparative statics. First and foremost, we find that an improvement in the baseline technology may backfire, by preventing AI supervision when it is actually desirable. Second, an improvement in the human ability to supervise AI can inefficiently reduce the equilibrium accuracy of the AI system (see Proposition 3).

It is worthwhile pointing out that our model could also be reinterpreted to fit the case of copyright infringements. Generative AI could infringe on copyright by composing texts, computer code, or music, or by creating images that do not sufficiently depart from existing, copyright-protected material. In an adaptation of our setup, the DM could be an agent who needs to compose some music or write some text or lines of a software code. In doing so, he can make use of generative AI, but may not know whether he is infringing on copyright. The DM would have to discard the output generated by the AI system if he discovers that the AI system used copyright-protected material. The results of our model can provide a rationale for business strategies currently pursued by AI developers. Recently, Microsoft announced that it will protect its paying customers who are sued for any copyright infringement over material generated by its AI software.[7] In the uncertain regulatory landscape, our model suggests that this strategy can foster users' adoption and even work as a commitment device for Microsoft to invest more to avoid copyright infringement.

[7]See The Financial Times article *"Microsoft pledges legal protection for AI-generated copyright breaches"* published on September 7, 2023.

1.1 Related Literature

The paper contributes to the emerging literature investigating the effect of liability on the adoption of AI. This topic has attracted the interest of both economists and legal scholars who have been looking into the optimal design of liability and regulation, also taking into account its potential impact on innovation.[8] Scholars have also adapted principal-agent models by exploring questions related to the delegation of authority to AI or a human agent (see Athey et al., 2020) or the impact of generative AI on strategic communication (see Gans, 2024b). In our model, an agent must decide whether to use and follow the AI for a decision that may cause harm. Similarly, Llanes and Madio (2024) analyzes AI firms' business strategies and regulatory interventions, such as transparency mandates and risk-management requirements, in a model where users may misperceive risks associated with the usage of AI systems. In our setup, while users do not misperceive risks, they do not directly observe the accuracy of the AI system, and their choices may harm others. While Dai and Singh (2025) studies how liability and insurance reimbursement affect physicians' decisions to use assistive AI in clinical care, we investigate how the allocation of liability between the potential AI operator and the AI developer affects the adoption of AI and its impact on the precision of the technology. Chen and Hua (2024b) examines liability for AI, focusing on a firm's choice of both a safety investment and the degree of AI autonomy (defined as user control). We take a different approach by investigating how liability influences the firm's investment in AI precision and a user's decision to monitor the accuracy of the information generated by the AI system

Although we primarily focus on mitigating hallucinations, as we mentioned in the introduction, our paper also contributes to the current debate on generative AI and copyright, which has lately been attracting scholarly attention. Gans (2024a) theoretically analyzes the issue of the fair-use standard, that is, whether to compensate creators whose content has been used to train generative AI models. Yang and Zhang (2024) develops a dynamic model to relate fair use and AI copyrightability (i.e., whether AI-generated

[8] Among the many contributions, see Galasso and Luo (2017, 2022), Guerra et al. (2022a,b), Guerreiro et al. (2023), Acemoglu and Lensman (2024), Buiten et al. (2023), and Buiten (2024). Gans (2025) examines recent research on the optimal rate of AI adoption in the presence of uncertainty about its potential harm.

content should enjoy copyright protection). Our work departs from other papers in this strand of the literature by focusing on the allocation of liability for copyright infringements between the AI developer and the AI operator.

Lastly, our paper is related to the economics literature studying how liability affects firms' incentives to learn about product risks and to invest in R&D activities to improve product safety.[9] This literature has focused on traditional products, whereas we consider the specificities of AI products and their interplay with human decisions. In our model, engaging in AI supervision enables the DM to acquire information about the reliability of the AI system and thus assess the expected harm a risky activity can cause. Alternative liability rules affect the DM's choice to supervise the AI and to act on what he learns. In this respect, there is a relationship to Shavell (1992). Both papers discuss the varying informational requirements that liability rules impose on the courts. We also consider how liability affects the AI developer's incentives to improve the reliability of the AI and its interplay with the DM's decision to collect information at a cost. Our paper also shares some similarities with Chen and Hua (2012), since both papers analyze the relationship between a firm's ex-ante investment and ex-post actions that may mitigate harm and how these activities are affected by liability. However, our model differs in two significant ways: first, we consider harm to both users and third parties, and second, we examine a context where these ex-post harm-mitigating actions are taken by users, whereas Chen and Hua (2012) focuses on firm-led responses. In this strand, a key predecessor to our work is Hay and Spier (2005) that studies the desirability of holding manufacturers liable for harms product users cause to third parties. As long as users are deep-pocketed and the manufacturers' investments in product safety are publicly observable, first best can be achieved by making users solely liable. However, making manufacturers liable for the shortfall not covered by users may be desirable. Hay and Spier (2005) does not explore a scenario in which a manufacturer's investment is not observed by users, which is the main focus of our paper. In our setup a trade-off arises because only liability concerns can motivate the AI developer to incur a cost and enhance the accuracy of the AI system.

[9]Recent contributions include Hua and Spier (2020), Henry et al. (2022), Chen and Hua (2024a), and Guadalupi et al. (2024).

2 Model

Consider an expert (e.g., a physician, an architect, or a lawyer) who must choose one of two activities $a \in \{s, x\}$, where s represents a safe activity (e.g., prescribing an established medication, selecting a proven design, or settling a case out of court, respectively) and x a risky activity (e.g., performing a surgery, developing an innovative design, or taking a case to trial, respectively). We normalize the expert's payoff from the safe activity s to 0. The risky activity x gives the expert a payoff $V > 0$. However, unlike the safe activity s, which causes no harm to anyone, activity x causes harm $H > V$ to a third party (e.g., a patient, a user, or a defendant, respectively) when the state of the world, denoted by θ, is bad. In addition, the expert suffers a disutility $d \geq 0$ when she chooses the risky activity and it turns out to be harmful. The parameter d captures the *reputational damage*, for example, the blow to a physician's reputation who performs an unsuccessful surgery, an architect's reputation whose innovative design fails, or a lawyer's reputation who loses a case during trial. The parameter d also captures the weight the expert attaches to the well-being of third parties who suffer harm because of her chosen activity.

The state is bad (i.e., $\theta = b$) with probability $q \in (0, 1)$, and is good ($\theta = g$) with probability $1 - q$. Therefore, parameter q captures the dangerousness of the risky activity. All players know the distribution of θ but not the realized state.

Before choosing which activity to carry out, the expert can make use of a generative AI system. The AI system generates a signal on θ, denoted s_{AI}, or, equivalently, a recommendation on the activity to be pursued, a, after analyzing and learning from existing datasets. The underlying data used to produce the signal may not be reliable, though. Formally, we say that the AI system (or, simply, the AI) is reliable, that is, $r_{AI} = r$, if the information used to generate the signal on the state of the world is exact. The AI system is unreliable, that is, $r_{AI} = u$, if the information used to generate s_{AI} is inexact. This latter scenario captures the artificial hallucinations AI systems may suffer from due to errors and deficiencies in their training data and their design.[10] We assume the

[10]For instance, the dataset used to train the generative AI system may contain false, incorrect, biased, outdated, or incomplete information, such as fake images, fraudulent financial data, or subjective statements. Generative AI models may then hallucinate if they fail to understand context or detect inaccuracies.

probability with which the AI is reliable, i.e., its *accuracy*, is $\rho \in [\rho_0, 1]$, where $\rho_0 \in [0, 1)$ is the baseline accuracy of the technology due to existing technological advancements freely available to the AI firm. Throughout, we assume that ρ_0 is publicly observed.

An AI firm (also referred to as an AI developer) can privately invest resources to increase the accuracy of the AI system from ρ_0 to ρ. For example, the AI firm can improve the quality of the dataset, employ multiple datasets, and conduct tests (e.g., the hold-out test) to improve the model's performance on unseen data.[11] We let $\Delta\rho := \rho - \rho_0$ and denote the cost of improving accuracy by $c(\Delta\rho, \rho_0)$, assuming that $c(0, \rho_0) = \frac{\partial c(0,\rho_0)}{\partial \Delta\rho} = 0$ for all ρ_0, $c(1 - \rho_0, \rho_0) = \infty$, $\frac{\partial c(\Delta\rho,\rho_0)}{\partial \Delta\rho} \geq 0$, $\frac{\partial^2 c(\Delta\rho,\rho_0)}{\partial \Delta\rho^2} > 0$, and that the cost function exhibits increasing differences in $(\Delta\rho, \rho_0)$, that is, $\frac{\partial^2 c(\Delta\rho,\rho_0)}{\partial \Delta\rho \partial \rho_0} > 0$. This last assumption captures the increasing difficulty the AI firm faces in raising AI accuracy when the baseline precision ρ_0 of the technology is higher. When the AI is reliable, the signal it generates is correct: that is, $s_{AI} = \theta$ if $r_{AI} = r$. Conversely, if the AI is unreliable, the signal is random: it produces an output based on inaccurate information that, by chance, can turn out to be correct.[12] In particular, we assume that $s_{AI} = g$ with probability $1/2$ and $s_{AI} = b$ with probability $1/2$ if $r_{AI} = u$, irrespective of the true state of the world. As a result, $Pr[s_{AI} = g|\theta = g] = Pr[s_{AI} = b|\theta = b] = \rho + \frac{1-\rho}{2}$; equivalently, the AI signal correctly identifies the state of the world θ with probability $\frac{1+\rho}{2}$.[13] The AI firm sets the price p for the use of its technology. This assumption is consistent with real-world observations: advanced AI tools, such as the latest versions of ChatGPT and Google's Gemini Advanced, often require paid subscriptions.

When the expert uses AI, she receives the signal s_{AI} and updates her belief about the state θ using the Bayes' rule. If the expert receives $s_{AI} = g$, she believes that the true

[11] For an informal overview of some of the recent techniques employed by AI developers to reduce hallucinations, see *"The 'hallucinations' that haunt AI: why chatbots struggle to tell the truth"*, published in the Financial Times on July 22, 2025.

[12] To provide an example, a hallucinating AI can generate a diagnosis that is correct, even though it is based on inexact information.

[13] In Appendix C, we show that our results would qualitatively hold if the number of states were larger than two and only one state were good.

state is $\theta = g$ with probability

$$\begin{aligned}\beta_g(\rho) =& Pr[\theta = g|s_{AI} = g] = \frac{Pr[s_{AI} = g|\theta = g]Pr(\theta = g)}{Pr[s_{AI} = g|\theta = g]Pr(\theta = g) + Pr[s_{AI} = g|\theta = b]Pr(\theta = b)} \\ =& \frac{(1+\rho)(1-q)}{(1+\rho)(1-q) + (1-\rho)q} \in [1-q, 1].\end{aligned}$$

Similarly, if the expert receives $s_{AI} = b$, she believes that the true state is $\theta = g$ with probability

$$\begin{aligned}\beta_b(\rho) =& Pr[\theta = g|s_{AI} = b] = \frac{Pr[s_{AI} = b|\theta = g]Pr(\theta = g)}{Pr[s_{AI} = b|\theta = g]Pr(\theta = g) + Pr[s_{AI} = b|\theta = b]Pr(\theta = b)} \\ =& \frac{(1-\rho)(1-q)}{(1-\rho)(1-q) + (1+\rho)q} \in [0, 1-q].\end{aligned}$$

The expert can engage in AI supervision by exerting effort to verify the exactness of the AI signal at a cost $k > 0$. We assume that AI supervision reveals a hallucination with probability $m \in (0, 1)$. The expert will not detect a hallucination when the AI is reliable (i.e., when there is no hallucination). Formally, let $s_m \in \{r, u\}$ be the signal produced from AI supervision. Then, $Pr[s_m = u|r_{AI} = u] = m$, and $Pr[s_m = u|r_{AI} = r] = 0$. Therefore, learning that the AI system is unreliable (i.e., $s_m = u$) reveals that the information the AI uses to generate the signal is inexact and the signal must consequently be discarded. As a result, $Pr[\theta = g|s_m = u] = 1-q$. By contrast, observing $s_m = r$ does not necessarily mean that the AI signal is indeed correct, because the expert's monitoring activity may fail to detect a hallucination.

If the expert engages in AI supervision, she further updates her belief about the state of the world θ. Suppose the expert receives $s_{AI} = g$, verifies its correctness, and finds the AI signal to be reliable (i.e., $s_m = r$). In this case, the expert updates her belief $\beta_g^r(\rho)$ about the state $\theta = g$ using the chain rule:

$$\begin{aligned}\beta_g^r(\rho) =& Pr[\theta = g|s_m = r \cap s_{AI} = g] = \frac{Pr[s_m = r|\theta = g \cap s_{AI} = g]Pr[\theta = g|s_{AI} = g]}{Pr[s_m = r|s_{AI} = g]} \\ =& \frac{\rho(1-q) + (1-m)\left(\frac{1-\rho}{2}\right)(1-q)}{\rho(1-q) + (1-m)\left(\frac{1-\rho}{2}\right)} \in [\beta_g(\rho), 1].\end{aligned}$$

However, if the expert finds the $s_{AI} = g$ signal to be unreliable (i.e., if $s_m = u$), the expert learns that the AI signal is worthless and cannot be used to update her belief $\beta_g^u(\rho)$ about

the state θ. Equivalently, $\beta_g^u(\rho) = 1 - q$.[14] Next, suppose that the expert engages in AI supervision after observing $s_{AI} = b$, and finds $s_m = r$. In this case, the expert updated belief $\beta_b^r(\rho)$ about the state $\theta = g$ is given by:

$$\begin{aligned}\beta_b^r(\rho) =& Pr[\theta = g | s_m = r \cap s_{AI} = b] = \frac{Pr[s_m = r | \theta = g \cap s_{AI} = b] Pr[\theta = g | s_{AI} = b]}{Pr[s_m = r | s_{AI} = b]} \\ =& \frac{(1-m)\left(\frac{1-\rho}{2}\right)(1-q)}{q\rho + (1-m)\left(\frac{1-\rho}{2}\right)} \in [0, \beta_b(\rho)].\end{aligned}$$

However, if the expert finds the $s_{AI} = b$ signal to be unreliable $s_m = u$, she understands that the AI signal cannot be used to update her belief about the state θ and must be discarded. Equivalently, $\beta_b^u(\rho) = q$.[15] As expected, an increase in the precision of AI monitoring m reduces uncertainty about the state θ: $\frac{\partial \beta_g^r(\rho)}{\partial m} > 0$ and $\frac{\partial \beta_b^r(\rho)}{\partial m} < 0$.

The sequence of events is as follows. First, nature draws the state of the world θ, which remains unknown to all the players. Second, the AI firm privately invests in AI accuracy ρ and sets the price p for the use of its technology. The expert observes the price p and forms a rational expectation of AI accuracy ρ. Next, the expert decides whether to purchase the AI system. If AI is purchased, the signal s_{AI} is generated and the expert decides whether to engage in AI supervision at a cost k. Subsequently, the expert chooses $a \in \{s, x\}$. Finally, payoffs are realized.

We compare alternative ways to assign liability for harm caused to third parties when the expert relies on AI during activity selection. We consider *AI-Operator liability* and *AI-Developer liability* regimes in which the expert and the AI firm are, respectively, held responsible for third-party damages. In addition, we also consider an *AI-Developer conditional liability* regime, in which the expert is liable unless she engages in AI supervision,

[14] It is straightforward that

$$\begin{aligned}\beta_g^u(\rho) = Pr[\theta = g | s_m = u \cap s_{AI} = g] =& \frac{Pr[s_m = u | \theta = g \cap s_{AI} = g] Pr[\theta = g | s_{AI} = g]}{Pr[s_m = u | s_{AI} = g]} \\ =& \frac{m\left(\frac{1-\rho}{2}\right)(1-q)}{m\left(\frac{1-\rho}{2}\right)(1-q) + m\left(\frac{1-\rho}{2}\right)q} = 1 - q.\end{aligned}$$

[15] Similar to $\beta_g^u(\rho)$, it is straightforward to see that

$$\begin{aligned}\beta_b^u(\rho) = Pr[\theta = b | s_m = u \cap s_{AI} = b] =& \frac{Pr[s_m = u | \theta = b \cap s_{AI} = b] Pr[\theta = b | s_{AI} = b]}{Pr[s_m = u | s_{AI} = b]} \\ =& \frac{m\left(\frac{1-\rho}{2}\right)q}{m\left(\frac{1-\rho}{2}\right)q + m\left(\frac{1-\rho}{2}\right)(1-q)} = q.\end{aligned}$$

in which case, liability is shifted to the AI firm. We examine the strengths and weaknesses of different liability regimes in inducing (1) the expert to select the welfare-maximizing activity a and (2) the AI firm to invest in the accuracy ρ of AI.

Next, we formally draw the distinction between high-risk and low-risk activities.[16] We refer to an activity a as a high-risk activity when $q > \hat{q} := \frac{V}{H+d}$, whereas an activity is defined as low risk if $q \leq \hat{q}$.[17] In most of the analysis, we focus on high-risk activities. We study low-risk activities in Section 5.2. Finally, we specify when reputational concerns are significant vs. minor. We say reputational concerns are significant when $d > \frac{V}{q}$.[18] Conversely, reputational concerns are minor when $d \in \left[\frac{V}{q} - H, \frac{V}{q}\right]$.

3 AI Supervision and Activity Selection

In this section, we study the incentives that AI-Operator liability and AI-Developer liability create for engaging in AI supervision and selecting the activity a for exogenous levels of p; that is, we focus on ex-post efficiency. We defer to Section 5.1 the analysis of AI-Developer Conditional liability.

Under AI-Operator liability, the DM is liable even when he uses and follows the AI system recommendation. Under AI-Developer liability, the AI developer is liable for the harm suffered by third parties if the AI system makes a recommendation that the DM follows and turns out to be incorrect. Clearly, if the DM does not use the AI system or does not follow its recommendation (i.e., the DM chooses $a = x$ even though $s_{AI} = b$), he would be liable for the harm such a decision causes.

Since the two regimes only differ with regard to who bears the compensation H when the DM follows the AI system's positive recommendation, we introduce the variable $\gamma \in \{0, 1\}$, where $\gamma = 1$ under AI-Operator liability, and $\gamma = 0$ under AI-Developer liability.

When the activities are high risk, under both liability regimes, observing $s_{AI} = b$ in stage 3 discourages the DM from undertaking $a = x$, whereas observing $s_{AI} = g$ improves

[16] Our distinction is somewhat reminiscent of the EU approach of regulating AI differently according to their riskiness. We refer the reader to the EU Artificial Intelligence Act.

[17] When based on the prior distribution of the state of the world (i.e., without knowing the true state of the world), pursuing an activity only if is low risk is socially desirable.

[18] When $d > \frac{V}{q}$, the fear of a reputational blow following harm to third parties affects the expert's choice under the AI-Developer liability regime.

the DM's belief that the state could be favorable. Thus, only in the latter instance would the DM contemplate pursuing $a = x$. However, only under AI-Developer liability can the DM shift liability for harm suffered to third parties to the AI developer when he follows a positive recommendation. Therefore, the DM may undertake $a = x$ even though his posterior belief that the state is favorable has not changed enough to justify the selection of a risky activity solely on efficiency grounds. It follows that social and private (i.e., the DM's) incentives for activity selection are not necessarily aligned under this regime. We now develop this intuition more formally.

3.1 Formal Analysis

Consider stage 3 and suppose $s_{AI} = g$. The DM must decide whether to double-check the AI signal. If he does not, he expects to get:

$$\max\{V - [1 - \beta_g(p)](\gamma H + d), 0\}.$$

The DM strictly prefers $a = x$ to $a = s$ when $p > p_A(\gamma)$, where[19]

$$p_A(\gamma) := \max\left\{\frac{q(\gamma H + d) - V}{q(\gamma H + d) - V + 2(1-q)V}, 0\right\}.$$

Note that $p_A(0) < p_A(1)$ for any $H > 0$. Moreover, $p_A(1) > 0$, whereas $p_A(0) = 0$ when $q \leq V/d$. That is, the DM is more likely to select the risky activity after observing a favorable AI signal under AI-Developer liability. If the DM does not engage in AI supervision after observing $s_{AI} = g$, his expected utility is:[20]

$$\pi_h^{nm}(g) = \begin{cases} V - [1 - \beta_g(p)](\gamma H + d), & \text{if } p > p_A(\gamma); \\ 0, & \text{if } p \leq p_A(\gamma). \end{cases}$$

If the DM engages in AI supervision after observing $s_{AI} = g$, he may or may not find out that the AI signal was a hallucination. If $s_m = u$, the DM learns that he cannot rely on the AI signal and he will pursue activity $a = s$ under AI-Operator liability. Under AI-Developer liability, that choice hinges on the verifiability of the result of AI supervision and qd. If the result of AI supervision is verifiable, the DM will not be able to shift liability to the AI developer following $s_m = u$ and will consequently choose $a = s$. If the result is

[19] Note that the numerator is always positive when $q > \hat{q}$ and $\gamma = 1$.

[20] The subscript h refers to high-risk activities, and the superscript nm stands for non-monitoring.

not verifiable, the DM could still shift liability to the AI developer, and he will select $a = x$ whenever $q < V/d$. If the DM learns the AI signal is reliable, that is, $s_m = r$, he believes that $\theta = g$ with probability $\beta_g^m(p)$ and selects $a = x$ if $V - [1 - \beta_g^m(p)](\gamma H + d) \geq 0$, that is, if $p \geq p_B(\gamma)$ where

$$p_B(\gamma) := \max\left\{\frac{(1-m)[q(\gamma H + d) - V]}{(1-m)[q(\gamma H + d) - V] + 2(1-q)V}, 0\right\}.$$

Note that $p_B(\gamma) = p_A(\gamma)$ when $m = 0$ and $p_B(\gamma)$ is decreasing in m. Moreover, $p_B(0) < p_B(1)$ for any $H > 0$, and the DM will always choose $a = x$ if AI supervision reveals that the AI system is reliable if $qd < V$ when $\gamma = 0$.

To determine the DM's incentive to engage in AI supervision, we first need to compute his expected utility from double-checking $s_{AI} = g$. We begin by considering the case in which the result of AI supervision is verifiable, which is relevant for AI-Developer liability. If s_m is verifiable, the DM will undertake the risky activity only if the AI supervision does not find evidence of a hallucination. Because monitoring is imperfect, the DM anticipates that with probability $1 - \beta_g(p)$, the AI system produced $s_{AI} = g$ when $\theta = b$ and AI supervision does not detect the hallucination with probability $1-m$. If so, by undertaking $a = x$, the DM will lose $\gamma H + d - V$. Moreover, the DM knows that even when $\theta = g$, AI supervision may reveal the favorable signal was in fact the product of an AI hallucination, in which case, he will refrain from undertaking $a = x$. This latter instance occurs with probability $Pr[s_m = u|\theta = g \cap s_{AI} = g] = \frac{m(1-p)}{1+p}$. As a result, the DM's expected utility from supervising the AI system when $s_{AI} = g$ is given by

$$\pi_h^m(g) = -k + \begin{cases} \beta_g(p)\left(1 - \frac{m(1-p)}{1+p}\right)V - [1 - \beta_g(p)](1-m)(\gamma H + d - V), & \text{if } p \geq p_B(\gamma); \\ 0, & \text{if } p < p_B(\gamma). \end{cases}$$

We note that the above expression is unaffected by the verifiability of the result of AI supervision. If reputational concerns are minor, that is, if $qd \leq V$, the DM will never engage in AI supervision and select $a = x$. When reputational concerns are significant, that is, if $qd > V$, the DM will not select $a = x$ after learning that the AI is unreliable, irrespective of whether the outcome of monitoring is verifiable.

We proceed to analyze the DM's incentive to engage in AI supervision after observing $s_{AI} = g$. If $p < p_B(\gamma)$ the DM never double-checks the AI signal: the DM will choose $a = s$ regardless of the outcome of AI supervision. For $p \geq p_B(\gamma)$, we need to contemplate

two cases. First is the scenario in which $p \in [p_B(\gamma), p_A(\gamma)]$. The relevant comparison is between

$$\beta_g(p)\left(1 - \frac{m(1-p)}{1+p}\right)V - [1-\beta_g(p)](1-m)(\gamma H + d - V) - k$$

and 0 because the DM selects $a = x$ only if AI supervision does not find that $s_{AI} = g$ is a hallucination. The DM weakly prefers to engage in AI supervision if

$$p \geq \underline{p}_h(\gamma) := \frac{(1-m)[q(\gamma H + d) - V] + k}{(1-m)[q(\gamma H + d) - V] + k + 2(1-q)(V-k)}.$$

Note this threshold is increasing in k and decreasing in m, and it holds that $\underline{p}_h(1) > \underline{p}_h(0)$. Moreover, $\underline{p}_h(\gamma)$ is also always higher than $p_B(\gamma)$ and is lower than $p_A(\gamma)$ if the following inequality is satisfied:[21]

$$\frac{m}{k} \geq \frac{q(\gamma H + d)}{V[q(\gamma H + d) - V]}. \tag{1}$$

Intuitively, the DM will supervise the AI system only if monitoring is relatively precise given its cost k. It is easy to see that Condition (1) is easier to satisfy when $\gamma = 1$ than when $\gamma = 0$ for any $H > 0$. The second scenario is the one in which $p \in [p_A(\gamma), 1]$. The relevant comparison is between

$$\beta_g(p)\left(1 - \frac{m(1-p)}{1+p}\right)V - [1-\beta_g(p)](1-m)(\gamma H + d - V) - k$$

and $V - [1-\beta_g(p)](\gamma H + d)$ because the DM selects $a = x$ when he does not engage in AI supervision. The DM weakly prefers to supervise the AI if

$$p \leq \overline{p}_h(\gamma) := \max\left\{\frac{m[q(\gamma H + d) - V] - k}{m[q(\gamma H + d) - V] - k + 2(1-q)k}, 0\right\}.$$

It is easy to see that $\overline{p}_h(\gamma) < 1$, whereas $\overline{p}_h(\gamma) > p_A(\gamma)$ if Condition (1) is satisfied. It also holds that $\overline{p}_h(1) > \overline{p}_h(0)$.

We can now write the DM's expected utility when $s_{AI} = g$. If AI monitoring is inefficient, that is, if Condition (1) is not satisfied, the DM never engages in AI supervision after a favorable AI signal, and his expected utility is $\pi_h^{nm}(g)$. If Condition (1) is satisfied, the DM's expected utility after observing $s_{AI} = g$ is

$$\pi_h^{DM}(g) = \begin{cases} 0, & \text{if } p < \underline{p}_h(\gamma); \\ \beta_g(p)\left(1 - \frac{m(1-p)}{1+p}\right)V - [1-\beta_g(p)](1-m)(\gamma H + d - V) - k, & \text{if } p \in [\underline{p}_h(\gamma), \overline{p}_h(\gamma)]; \\ V - [1-\beta_g(p)](\gamma H + d), & \text{if } p > \overline{p}_h(\gamma). \end{cases}$$

[21] When $\gamma = 0$, it must also be that $qd > V$.

Thus, the DM will proceed to select $a = x$ after observing $s_{AI} = g$ without double-checking the AI signal only if the precision of the AI signal is high enough; for intermediate values of AI precision, the DM will double-check the favorable AI signal and select $a = x$ only if monitoring does not reveal that s_{AI} was generated by a hallucination. If the AI is not precise enough, the DM will not undertake the risky activity. More efficient AI supervision, measured by $\frac{m}{k}$, increases the region of parameters in which AI supervision takes place.

AI adoption. To analyze the DM's decision to adopt AI, we need to specify his expected utility from using the AI system. Suppose that Condition (1) holds and, if $\gamma = 0$, it must also be that $q > V/d$. We can describe the expected utility of the DM in three different parameter regions as follows:

$$\pi_h^{DM} = -T \tag{2}$$

$$+ \begin{cases} 0, & \text{if } p < \underline{p}_h(\gamma). \\ (1-q)\left[p + \frac{1-p}{2}(1-m)\right]V - \frac{q(1-p)(1-m)}{2}(\gamma H + d - V) - \frac{(1-q)(1+p)+q(1-p)}{2}k, & \text{if } p \in \left[\underline{p}_h(\gamma), \overline{p}_h(\gamma)\right]; \\ \frac{(1-q)(1+p)}{2}V - \frac{q(1-p)}{2}(\gamma H + d - V), & \text{if } p > \overline{p}_h(\gamma). \end{cases}$$

When the activity is high risk, the DM will undertake $a = x$ only if the AI signal is favorable and sufficiently precise. For intermediate values of p, the DM will also engage in AI supervision to be more confident about the reliability of the recommendation received. For high values of AI precision, the DM will choose $a = x$ without double-checking a favorable AI signal.

If Condition (1) is not satisfied and, when $\gamma = 0$, it holds that $q > V/d$, AI supervision never takes place and the DM's expected utility from using AI is

$$\pi_h^{DM} = -T + \begin{cases} 0, & \text{if } p \leq p_A(\gamma). \\ \frac{(1-q)(1+p)}{2}V - \frac{q(1-p)}{2}(\gamma H + d - V), & \text{if } p > p_A(\gamma). \end{cases} \tag{3}$$

If $\gamma = 0$ and $q \in [\hat{q}, V/d]$, the DM's expected utility is

$$\pi_h^{DM} = -T + \frac{(1-q)(1+p)}{2}V - \frac{q(1-p)}{2}(d - V) \tag{4}$$

because $p_A(0) = 0$.

3.2 Comparison of Liability Regimes

In this subsection, we aim to compare the incentives that the two distinct liability regimes analyzed above create for AI supervision and adoption and, ultimately, for the selection of the most efficient activity a. We begin by highlighting the following result.

Remark 1. *The DM's incentives for AI supervision and for activity selection are socially efficient under AI-Operator liability.*

The observation that AI-Operator liability is ex-post efficient facilitates the comparison between the two regimes: any difference in AI supervision and activity-selection incentives can be viewed as a distortion engendered by assigning liability to the AI developer.

Performing this comparison according to two criteria will be useful. The first criterion is that of *supervision efficiency*. We note and henceforth will say that performing AI supervision is *socially efficient* when Condition (1) is satisfied for $\gamma = 1$. As observed earlier, Condition (1) is more stringent when $\gamma = 0$, and, as a result, there exist parameter values for which AI supervision would occur only under AI-Operator liability. The second criterion we employ is that of reputational relevance.

Efficient supervision and significant reputational concerns. Although the finding that allocating liability to the AI developer distorts incentives for AI supervision and for the selection of the activity may appear intuitive, what is remarkable is that AI-Developer liability may lead to excessive supervision of the AI system or even to the suboptimal selection of safe activities. The next proposition illustrates when these counterintuitive results may occur.

Proposition 1. *Let $q > V/d$ and AI supervision be socially efficient. AI-Developer liability distorts the incentives to supervise AI and to select the efficient activity. Although AI-Developer liability may entail too little AI supervision and excessive selection of the risky activity, this liability regime may also lead to*

(a) excessive AI supervision when:

$$\frac{m}{k} \geq \frac{qd}{V(qd-V)};$$

(b) too little risk-taking otherwise.

If Condition (1) is satisfied for $\gamma = 0$, AI supervision takes place under both legal environments. AI-Developer liability leads to over-supervision of AI when $p \in [\underline{p}_h(0), \underline{p}_h(1)]$: the ability to shift liability to the AI developer makes the DM less cautious. However, in this interval, given that the expected reputational damage d can be significant and the monitoring technology is efficient, the DM engages in AI supervision after observing $s_{AI} = g$. Conversely, if $p \in [\overline{p}_h(0), \overline{p}_h(1)]$, AI-Developer liability is associated with too little supervision: the observation of $s_{AI} = g$ is enough to induce the DM to select the risky activity $a = x$. These observations imply that activity selection will be distorted with AI-Developer liability. This case is illustrated in Figure 1.

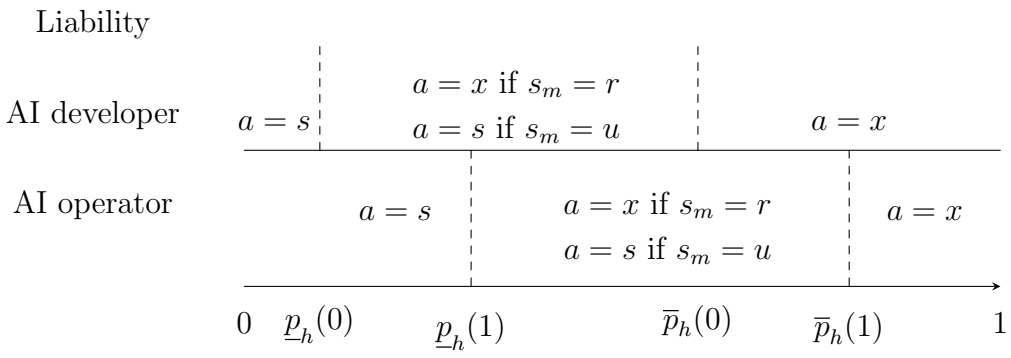

Figure 1: **Significant reputational concerns (a)**, that is, $q > \frac{V}{d}$. This figure illustrates the DM's activity choice under different liability regimes (on the y-axis) for different levels of p (on the x-axis). In drawing this figure, we assume that Condition (1) holds for $\gamma = 0$.

If Condition (1) is not satisfied for $\gamma = 0$, AI supervision may take place only if the DM retains liability for harm. Thus, AI-Developer liability results in under-supervision and may cause either too little or too much risk-taking, depending on whether $p_A(0)$ is at the right or at the left, respectively, of $\underline{p}_h(1)$. The intuition for why too little selection of the risky activity with AI-Developer liability may occur is the following. If p is low enough, the DM anticipates that he will not double-check a favorable signal, and since reputational concerns are meaningful, he prefers not to use the AI and selects the safe activity. Conversely, under AI-Operator liability, for the same signal precision, the DM

is willing to use the AI system, knowing that he will have an incentive to monitor the AI output before selecting the activity. This case is illustrated in Figure 2.

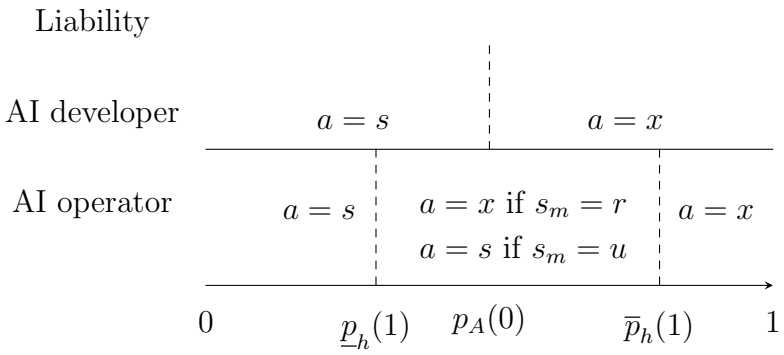

Figure 2: **Significant reputational concerns (b)**, that is, $q > \frac{V}{d}$. This figure illustrates the DM's activity choice under different liability regimes (on the y-axis) for different levels of p (on the x-axis). In drawing this figure, we assume that Condition (1) holds for $\gamma = 1$ but not for $\gamma = 0$.

Inefficient supervision and significant reputational concerns. AI supervision does not occur under either legal environment. Because $p_A(0) < p_A(1)$, we can conclude that AI-Developer liability results in excessive selection of the risky activity. In particular, when $p \in [p_A(0), p_A(1)]$, $a = x$ can occur only if the AI developer is liable for the harm suffered by third parties: when deciding whether to undertake the risky activity, the DM becomes more reckless when he knows that he can shift liability to the AI developer.

Minor reputational concerns. When the AI developer is liable, the DM always undertakes $a = x$ after $s_{AI} = g$, without bothering to double-check the signal produced by the AI system. Clearly, the incentives to undertake the efficient activity are greatly distorted. We report this observation in the following remark.

Remark 2. *Let $q \in [\hat{q}, V/d]$. AI-Developer liability entails too little supervision and excessive selection of the risky activity.*

Figure 3 helps illustrate the issue of excessive risk-taking that may arise under AI-Developer liability when reputational concerns are minor.

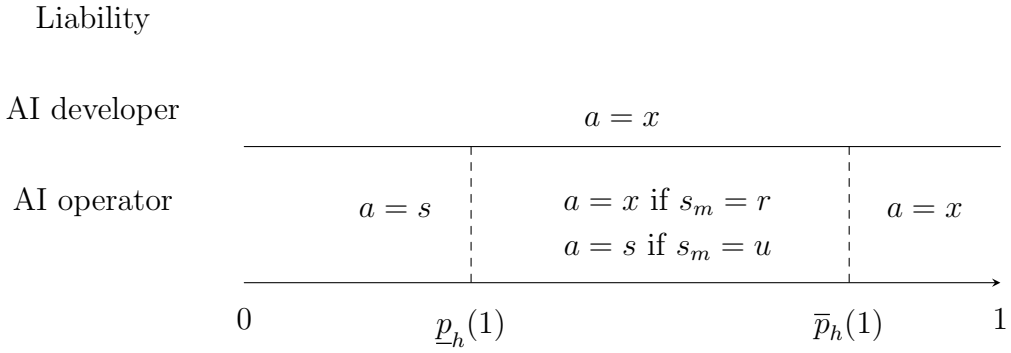

Figure 3: **Liability concerns and minor reputational concerns**, that is, $q \in [\hat{q}, \frac{V}{d}]$. This figure illustrates the DM's activity choice under different liability regimes (on the y-axis) for different levels of p (on the x-axis). In drawing this figure, we assume that Condition (1) holds.

4 AI Developer's Investment and Industry Profits

We now shift focus to ex-ante efficiency; in particular, we proceed to examine the AI developer's investment and pricing decisions.

The AI developer's expected utility at stage 2 is given by

$$\Pi = T - c(\Delta p, p_0) - (1 - \gamma) Pr[H] H,$$

where $Pr[H]$ is the probability that harm occurs. The AI developer chooses $T \geq 0$ and $p \in [p_0, 1]$ to maximize Π subject to the AI operator's acceptance of the contract.

As p is privately observed by the AI developer and the DM is only aware of the baseline accuracy of the technology p_0, assigning liability to the AI operator may cause under-investment, as we point out in the next remark, where we define p^{DM} as the actual accuracy choice made by the AI developer under AI-operator liability.

Remark 3. *Assigning liability to the AI operator leads to* $p^{DM} = p_0 = p^e$.

Intuitively, since p is privately observed, the AI developer cannot invest in the accuracy of the technology to directly demand a higher payment from the DM. Her only incentive to increase p stems from the willingness to avoid the expected damages she would have to pay in the case in which harm occurs. This incentive is absent if she is shielded from liability, though.[22] Conversely, she will have an incentive to boost the accuracy of the technology under AI-Developer liability.

[22] In our model, under neither liability regime can the AI developer signal a higher accuracy of the AI system by charging a higher price. Although some authors have shown that, in some contexts, the price

Therefore, the private observability of the AI accuracy (or, equivalently, of the AI developer's investment in the accuracy of the technology) gives rise to a trade-off. On the one hand, AI-Operator liability ensures ex-post efficiency; on the other hand, this liability regime entails no investment in the accuracy of the technology. The need to stimulate investment in the accuracy of the AI system may then make AI-Developer liability desirable. Under this alternative liability regime, the AI developer has an incentive to improve the accuracy of the technology p, being aware that she will be held accountable for the harm that the AI system may cause to third parties. Note that, because the DM will anticipate that the AI developer may invest in the accuracy of the AI system, he will be willing to pay a higher price for the use of the technology. Yet, the investment can only affect the price indirectly, by changing the DM's equilibrium expectation of p. For this reason, namely, because in choosing p the AI developer will overlook its direct impact on the DM's expected utility, her investment decision will fall short of ensuring social efficiency. We formalize these observations below. First, we show that three values of p are possible in equilibrium and we let p^{AI} denote the equilibrium choice of p under AI-Developer liability.

Lemma 1. *Under AI-Developer liability,* $p^{AI} \in \{p_0, p^m, p^{nm}\}$*, with* $p^{nm} := p_0 + \Delta p^{nm}$ *and* $p^m := p_0 + \Delta p^m$*, where* Δp^m *and* Δp^{nm} *are uniquely defined by:*

$$\frac{\partial c(\Delta p^m, p_0)}{\partial \Delta p} = \left(\frac{(1-m)qH}{2}\right); \quad \textit{and} \quad \frac{\partial c(\Delta p^{nm}, p_0)}{\partial \Delta p} = \left(\frac{qH}{2}\right),$$

respectively. Note that for any p_0*,* $\Delta p^{nm} > \Delta p^m > 0$*. The DM will correctly anticipate* $p^e = p^{AI}$*.*

The AI developer will not improve the signal precision when she anticipates that the AI system will not be used, and will invest either Δp^m or Δp^{nm} otherwise. The former (Δp^m) is the increase in the AI system precision when the AI developer expects the DM to supervise the AI system and the latter (Δp^{nm}) when she expects the DM not to engage in AI supervision. The equilibrium accuracy levels must be compatible with the DM's incentive to (not) engage in AI supervision when $s_{AI} = g$. Having characterized the

can be used to signal quality (e.g., see Daughety and Reinganum, 1995, 2008), this mechanism generally depends on higher quality (i.e., a more accurate AI system in our setting) having a higher marginal production cost.

equilibrium investment in the AI system accuracy under AI-Developer liability, we can determine the optimal transfers and compare industry profits under the two alternative liability regimes. The optimal transfers when $p^{AI} = p^{nm}$ and $p^{AI} = p^{m}$ are given by:

$$\begin{aligned} T(p^{nm}) =& \frac{(1-q)(1+p^{nm})}{2} V - \frac{q(1-p^{nm})}{2}(d-V), \\ T(p^{m}) =& (1-q)\left[p^{m} + \frac{1-p^{m}}{2}(1-m)\right] V - \frac{q(1-p^{m})(1-m)}{2}(d-V) \\ & - \frac{(1-q)(1+p^{m}) + q(1-p^{m})}{2} k, \end{aligned}$$

respectively.

We proceed to define industry profits as the sum of the DM's and the AI developer's expected profits.[23] As we highlight in the next proposition, AI-Operator liability can dominate only when it encourages AI supervision.

Proposition 2. *Industry profits can be strictly higher with AI-Operator liability only if AI supervision is socially efficient and* $p_0 \in [\underline{p}_h(1), \overline{p}_h(1)]$. *Moreover, there exists* $\tilde{p}_0 < \overline{p}_h(1)$ *such that for any* $p_0 > \tilde{p}_0$, *AI-Developer liability leads to strictly higher industry profits.*

These results may appear surprising because they stand in stark contrast to the findings of Section 3.2, which has shown that AI-Developer liability entails distorted choices once the AI system is adopted. Yet, because the AI developer can always decide to set such a high price for the technology that the DM refuses to purchase it, the AI system will not be used if the AI developer anticipates that the DM's choices will be reckless. Moreover, AI-Developer liability spurs investment in the precision of the technology, which creates more surplus that can be shared between the AI developer and the DM. AI-Operator liability turns out to be profitable only if it induces the DM to double-check the AI signal. Although the AI precision may be lower, the beneficial effect of supervising the AI system may more than compensate for it.

The second part of the proposition highlights a non-trivial relationship between the baseline precision of the technology, p_0, and the most profitable liability regime. When p_0 is very low, neither regime will lead to technology adoption, because the activity is high-risk. When p_0 takes intermediate values, AI-Operator liability dominates under the

[23] Note that as long as the negative externality H is entirely taken into account under fully compensatory damages, industry profits are also a measure of social welfare.

conditions described in the previous paragraph. However, for p_0 sufficiently high, AI-Developer liability always proves superior. To understand why, consider that for p_0 high enough, no AI supervision occurs under either AI Operator or AI-Developer liability, and the latter is the only one that stimulates investment to further the precision of the AI system. If we interpret the baseline precision of the AI system as the quality of the current technology, which evolves depending on factors external to the interaction studied in our model (e.g., research carried out in universities or research centers, breakthroughs obtained by firms operating in the technological sector), our results suggest that the optimal liability regime may change over time. More specifically, when the state of the technology is not too advanced, AI-Operator liability may be preferred to encourage adoption and much-needed AI supervision, whereas when the state of the technology is sufficiently advanced as to make AI supervision inessential, AI-Developer liability is superior because it stimulates investment to further improve the accuracy of AI.

We provide a graphical representation of the relationship between baseline precision of the technology, p_0, and industry profits in Figure 4. There, industry profits under AI-Operator liability are represented by a dashed blue line, whereas those under AI-Developer liability are represented by a solid red line. In the figure, AI-Operator liability is the only regime that induces AI supervision and, for p_0 sufficiently small, that is, for $p_0 \in [\underline{p}_h(1), \tilde{p}_0)$, it leads to higher industry profits. Conversely, AI-Developer liability maximizes industry profits for p_0 high enough, that is, for $p_0 > \tilde{p}_0$.

We conclude this section by providing the most intriguing results of the comparative statics analysis.[24] First, when d is relatively small, the AI developer has a strong incentive to invest in p because the DM would always select $a = x$ after observing $s_{AI} = g$, without double-checking its reliability. When d is relatively high, the DM will not follow the AI signal unless it is precise enough. If the achievement of that minimum level of accuracy is not compatible with the AI developer's incentives, which are known to the DM, she will refrain from investing. It follows that an increase in reputational concerns may lead to lower accuracy of the AI signal. Second, superior AI supervision ability, as measured by

[24] We provide additional comparative statics in Appendix B.

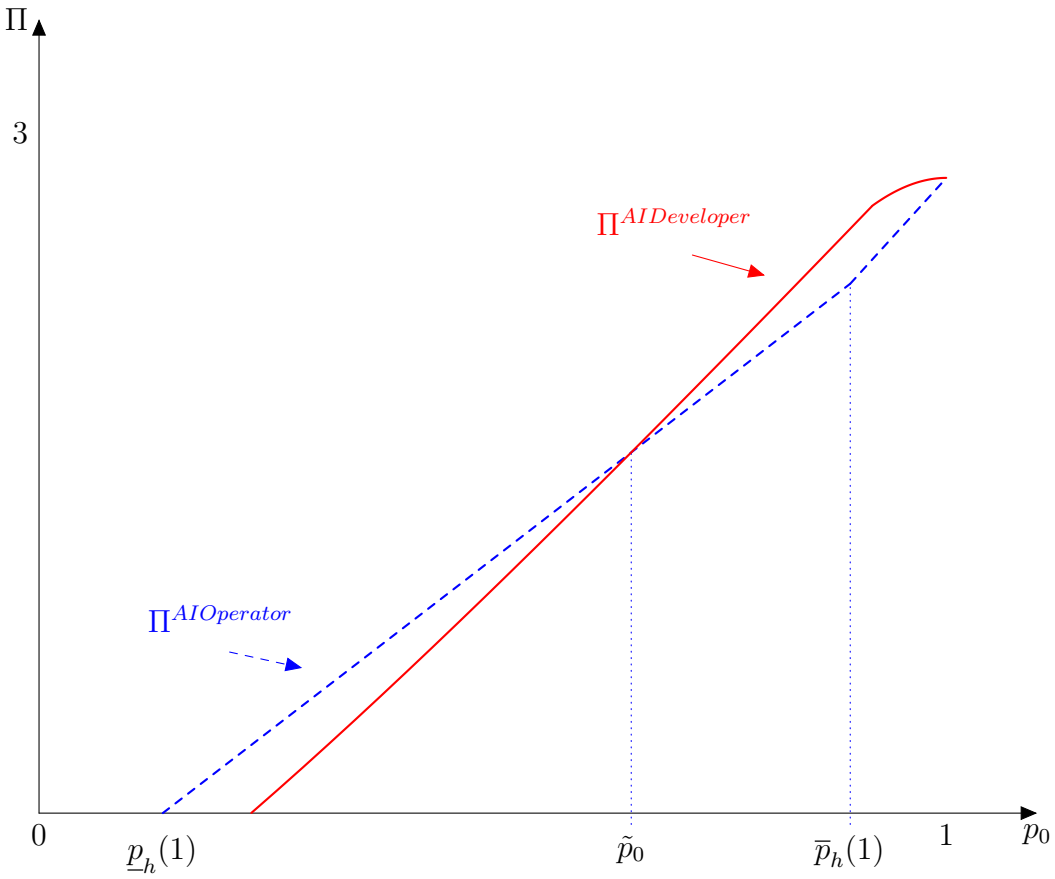

Figure 4: Industry profits. The figure illustrates industry profits as function of p_0 under the two liability regimes. We assume the following cost function and parameter values: $\frac{c[\Delta p(1+p_0)]^2}{2}$, $V = 4$, $d = 4$, $H = 20$, $q = 0.3$, $m = 0.8$, $k = 0.2$, $c = 10$. Accordingly, minor reputational concerns exist and Condition (1) is satisfied for $\gamma = 1$.

a higher m, may actually backfire by inducing the AI developer to choose an inefficiently low level of AI precision. Third, an increase in the baseline precision of the AI system p_0 may result in lower industry profits. These findings are described in the next proposition.

Proposition 3. *An increase in reputational concerns d or a higher AI supervision ability m may inefficiently lead to a reduction in the accuracy of the AI system p. A higher baseline precision p_0 may reduce industry profits.*

Notably, although lowering p when m takes a higher value may be efficient, here we are referring to an inefficiently low choice of AI signal precision due to the AI developer not internalizing the DM's profit when making her own investment. To understand why, suppose that $q > V/d$, Condition (1) is satisfied for $\gamma = 0$, and $p^{nm} > \overline{p}_h(0)$ whereas $p^m \in [\underline{p}_h(0), \overline{p}_h(0)]$. The AI developer weakly prefers p^{nm} to p^m if

$$\underbrace{\frac{qH}{2}[(1-p^m)(1-m)-(1-p^{nm})]}_{\Delta Liability} \geq \underbrace{c(p^{nm}-p_0)-c(p^m-p_0)}_{\Delta c}. \tag{5}$$

That is, the reduction in expected liability must at least offset the higher investment cost. The payment the DM receives does not affect the sign of this preference relation. An increase in m increases the likelihood that the above inequality does not hold. However, it could be that (5) is no longer satisfied following the increase in m, whereas it would still be efficient to choose $p = p^{nm}$ if

$$T(p^{nm}) - c(p^{nm}-p_0) - \frac{q(1-p^{nm})H}{2} > T(p^m) - c(p^m-p_0) - \frac{q(1-p^m)(1-m)H}{2}.$$

Consider now the effect of a change in d. Suppose that reputational concerns are initially minor and the AI developer finds choosing $p^{AI} = p^{nm}$ to be profitable. If reputational concerns increase and become meaningful, the AI developer may set $p^{AI} = p_0$ even if choosing $p > p_0$ could still be socially optimal.

Lastly, we briefly provide the intuition for why industry profits may decrease when p_0 takes a higher value, which occurs under AI-Developer liability and owes to the non-observability of p^{AI}: an increase in the baseline precision of the AI system may make $p^m > \overline{p}_h(0)$ and force the AI developer to choose $p^{AI} = p^{nm}$ even though doing so reduces profits.[25]

[25] This decrease in profits may make AI-Operator liability preferable when $p_0 \in (\overline{p}_h(0) + \Delta p^m, \overline{p}_h(1)]$. Yet, a drop in industry profits may occur following a shift from AI-Developer liability with p_0 below $\overline{p}_h(0) + \Delta p^m$ to AI-Operator liability with p_0 above $\overline{p}_h(0) + \Delta p^m$.

5 Extensions

In this section, we develop several extensions of the analysis carried out so far, which will also enable us to shed light on the role of some of the modeling assumptions. Some parts of the formal analyses, as well as the technical proofs, are relegated to Appendix B.

5.1 AI-Developer Conditional liability

We now consider an alternative liability regime whereby the AI operator is liable if he does not double-check the recommendation provided by the AI. If he does and some harm occurs, the AI developer will be on the hook. Relative to AI-Developer liability, this regime, which we refer to as *AI-Developer Conditional liability*, strengthens the DM's incentive to supervise the AI system and leads to socially excessive AI supervision. In turn, it affects the AI Developer's investment in the precision of the technology and the incentives to adopt the AI system. The formal analysis is fully developed in Appendix B. Here, we point out the key differences that arise with the liability regimes previously analyzed.

AI supervision and activity selection. The DM's incentive to engage in AI supervision will be critically affected by (i) the possibility of *verifying the result of AI supervision*, that is, whether the DM observed $s_m = r$ or $s_m = u$, and (ii) the relevance of reputational concerns d. Specifically, when either the result of AI supervision is verifiable or reputational concerns are significant, that is, when $q > \frac{V}{d}$, the DM wants to pursue $a = s$ if he learns the AI system was unreliable. We find that a necessary condition for the DM to engage in AI supervision under this regime is that

$$m \geq \frac{qdk}{V[qd - V]} - \frac{qH(V - k)}{V[qd - V]}. \tag{6}$$

In this case, the DM will double-check the AI signal when $p \in [\underline{p}_h(0), \overline{p}_h^{AIC}]$, where

$$\overline{p}_h^{AIC} := \frac{m(qd - V) - k + qH}{m(qd - V) - k + qH + 2(1 - q)k}.$$

Notably, compared with both AI Operator and AI-Developer liability, the scope of AI supervision is wider under this alternative legal regime for two reasons. First, supervision is more likely to be incentive compatible, because Condition (6) can be more easily satisfied

than Condition (1) for $\gamma = 1$. Second, whenever it is incentive compatible, AI supervision takes place more often as $\overline{p}_h^{AIC} > \overline{p}_h(1)$. Ultimately, the reason lies in the ability of shifting liability to the AI developer only conditional on supervising the AI result. This result is illustrated in Figure 5. Note, when $p \in (\overline{p}_h(1), \overline{p}_h^{AIC})$, AI Conditional liability may imply too little risk-taking, because the DM engages in AI supervision and may end up choosing $a = s$.

Consider now the scenario in which the result of AI supervision is not verifiable and reputational concerns are minor, that is, $q \in [\hat{q}, V/d]$. This liability regime creates incentives to engage in AI supervision for mere liability shifting purposes. We find that AI supervision will take place when $p \in [\underline{p}_{hL}^{AIC}, \overline{p}_{hL}^{AIC}]$, where two new threshold values are given by[26]

$$\underline{p}_{hL}^{AIC} := \frac{V - qd - k}{V - qd - k - 2(1-q)(V-k)}, \quad \overline{p}_{hL}^{AIC} := \frac{-qH + k}{-qH + k - 2(1-q)k}.$$

With respect to the other liability regimes, the DM's incentive to supervise is strengthened. Accordingly, we find that $\underline{p}_{hL}^{AIC} < \underline{p}_h(1)$ and $\overline{p}_{hL}^{AIC} > \overline{p}_h(1)$. This result is illustrated in Figure 6.

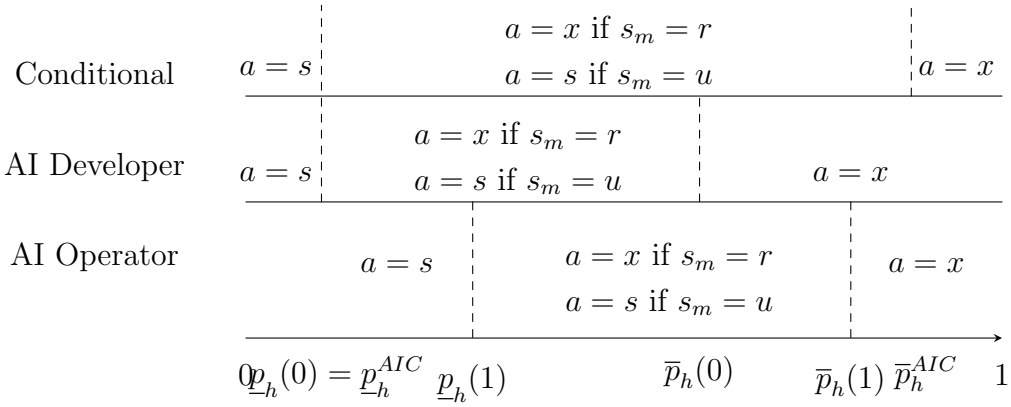

Figure 5: **Significant reputational concerns**, that is, $q > \frac{V}{d}$. This figure illustrates the DM's activity choice under different liability regimes (on the y-axis) for different levels of p (on the x-axis). In drawing this figure, we assume Condition (1) holds for $\gamma = 0$ and the result of AI supervision is verifiable.

[26] We use the subscript L to refer to this special case in which reputational concerns are minor.

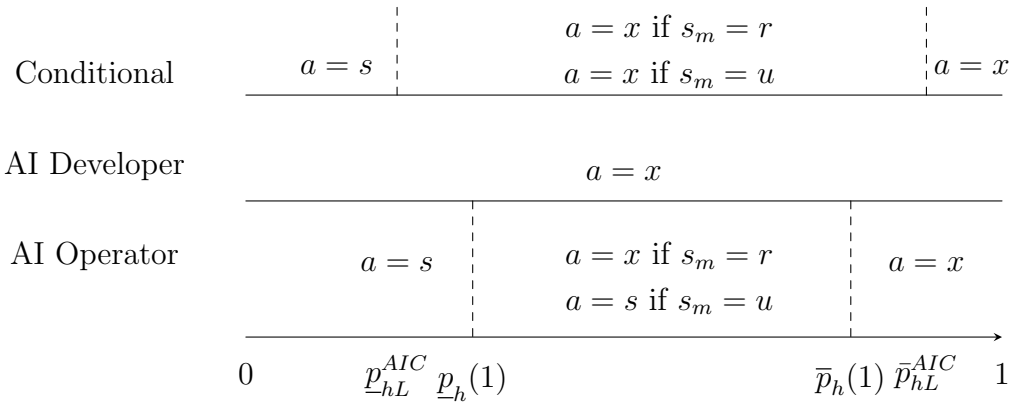

Figure 6: **Minor reputational concerns**, that is, $q \in \left[\hat{q}, \frac{V}{d}\right]$. This figure illustrates the DM's activity choice under different liability regimes (on the y-axis) for different levels of p (on the x-axis). In drawing this figure, we assume Condition (1) holds and the result of AI supervision is not verifiable.

We summarize the above findings in the following remark.

Remark 4. *AI-Developer Conditional liability results in excessive AI supervision relative to the other liability regimes and may lead to either excessive or too little risk-taking.*

AI developer's investment. To identify which liability regime maximizes industry profits, we focus on the case in which p is not observable so that the only driver of investment is the willingness to reduce the expected liability burden. With AI-Developer Conditional liability, the AI developer is liable only when the DM engages in AI supervision, which may occur in equilibrium only for intermediate levels of p_0. When the baseline precision of the technology is sufficiently high, the DM will not double-check the AI recommendation and the AI developer will be off the hook and have no incentive to invest. As a result, expected industry profits will coincide with those obtained under AI-Operator liability and AI-Developer liability will always be preferred. For intermediate levels of p_0 the case for AI-Operator liability weakens somewhat relative to that described in Proposition 2, because AI-Developer Conditional liability can lead to efficient decisions ex post without compromising AI precision. The following proposition characterizes the conditions for AI-Operator liability to dominate.

Proposition 4. *If engaging in AI supervision is verifiable, industry profits can be strictly higher with AI-Operator liability only if the following conditions simultaneously hold:*

(i) Condition (1) is satisfied for $\gamma = 1$;

(ii) $p_0 \in [\underline{p}_h(1), \overline{p}_h(1)]$;

(iii) reputational concerns are minor, i.e., $q \in \left[\hat{q}, \frac{V}{d}\right]$;

(iv) the result of AI supervision is not verifiable.

Moreover, there exists $\check{p}_0 < \overline{p}_h(1)$ *such that for any* $p_0 > \check{p}_0$*, AI-Developer liability leads to strictly higher industry profits.*

Although the above proposition may seem to dramatically reduce the scope for AI-Operator liability, we must stress that some of the conditions may not be hard to meet. First and foremost, under AI-Developer Conditional liability, to escape responsibility, the DM must be able to prove that he engaged in AI supervision. Even so, unless reputational concerns are meaningful, the DM must also be able to show that he did not find evidence that the AI system was unreliable. Thus, such a negligence regime may not always be implemented and, even when feasible, the litigation costs associated with its proper functioning may be unreasonably high.

5.2 Further Extensions

Below, we explore the implications of some of the assumptions of our model and we extend the model in several directions.

Public observability of the AI accuracy. A key assumption of our model is that the AI developer privately observes the accuracy of the technology. If p is *publicly observed*, because of Remark 1 and the AI developer's ability to fully extract the DM's expected surplus, AI-Operator liability leads to an optimal investment decision and, therefore, this regime achieves first-best, as we observed in the following remark.

Remark 5. *If* p *is publicly observable, assigning liability to the AI operator leads to efficient investment in the accuracy of the technology ex-ante and efficient activity selection by the DM ex-post.*

Remark 5 shows that the equilibrium choice of p under AI-Operator liability is the first best. Notably, there is a link between Remark 5 and Proposition 1 in Hay and Spier (2005): despite some modeling differences, the intuition for why private and social incentives

are aligned is similar.[27] Hay and Spier (2005) extensively explores the implications of the user's limited financial resources for the optimal design of liability, whereas we have considered the scenario in which the AI developer's investment is her private information. This assumption appears realistic given that the specifics of an AI system's training data and architecture are often proprietary information and thus not disclosed to the public, who can then only be aware of the generic precision of an off-the-shelf model, which is what we have referred to as the baseline precision of the technology, p_0.

No-liability benchmark. In the absence of liability, that is, if neither the AI developer nor the DM would have to compensate the harm incurred by third parties, the players' choices would combine the worst outcomes of the two regimes we have examined. Specifically, the DM's decision to engage in AI supervision and select the risky activity would only be constrained by the prospect of reputational damage d, thereby replicating the suboptimal behavior associated with AI-Developer liability. At the same time, as the AI developer would be exempt from liability, she would have no incentive to invest in improving the AI system accuracy and, as a consequence, the investment would be as poor as under AI-Operator liability. This highlights the need for a liability regime to improve the strategic choices made by either the AI developer or the DM.

Harm on DM. We have assumed that the DM faces an irrecoverable reputational damage d if things go awry, whereas third parties (e.g., innocent bystanders) incur a separate harm H. The results of the model remain unchanged if the DM bears a fraction of H.[28] Under AI-Operator liability, alongside bearing his own direct harm and reputational loss, the DM would be held accountable for compensating the other victims for their incurred damages. Conversely, under AI-Developer liability, the AI developer would have to fully compensate the DM as well as the other affected parties if an adverse outcome

[27] In Hay and Spier (2005), the product market is perfectly competitive, whereas our model considers a monopolistic provider of the AI system. Therefore, in our model, there is no quantity distortion. However, because there is a (representative) user with a unit demand, the AI developer can fully extract his surplus.

[28] A plausible example where the DM, rather than third parties, directly incurs harm occurs when a firm deploys an AI system to develop a proprietary internal software, and this results in an adverse outcome, like the loss of sensitive data.

occurs. Under either liability regime, the DM's and the AI developer's payoff functions are the same as those presented in the previous sections. However, it is worth remarking that the necessity for a liability regime varies with the identity of the harm-bearer. To understand why, consider that, without liability, the DM's decisions would be closer to the first-best the higher his share of the harm H. In the limit case where the DM bears the entire harm, the observed outcome would be indistinguishable from that under AI-Operator liability, thereby negating any additional benefit from such a formal liability assignment.

Low-Risk Activities. Here, we shift focus to low-risk activities, that is, activities for which $q \leq \hat{q}$. A prominent difference from what we have highlighted in the analysis of high-risk activities is that now the DM's activity and AI supervision choices are the same under AI Operator and AI-Developer liability. When $s_{AI} = g$, the DM always selects $a = x$ irrespective of the AI signal precision and of the allocation of liability without monitoring the AI system. When $s_{AI} = b$, the DM may decide to pursue $a = x$. Before making this choice, he may find double-checking the AI signal beneficial. However, the DM cannot shift liability to the AI developer if he chooses $a = x$ and things go awry: indeed, the AI was correct in discouraging the DM from undertaking the risky activity. Hence, there is no distortion in the choice of the activity under AI-Developer liability. As for AI adoption and investment incentives, consider that if p is privately observable, the AI developer will have an incentive to invest in the accuracy of the AI system only if she can be held liable. We summarize these observations in the following proposition:

Proposition 5. *Both AI Operator and AI-Developer liability are ex-post efficient, but only the latter regime provides incentives to invest in AI accuracy when p is privately observable. Hence, AI-Developer liability always leads to higher industry profits than AI-Operator liability.*

Also note that AI-Developer Conditional liability distorts the DM's incentives for AI supervision and activity selection. For illustration, consider that if the activity is low risk and $s_{AI} = g$, the DM will always pursue $a = x$. However, he may decide to engage in AI supervision only to shift liability to the AI developer. Formally, the DM's expected

utility if he does not engage in AI supervision after observing $s_{AI} = g$ is

$$V - [1 - \beta_g(p)](H + d).$$

If he engages in AI supervision, his expected utility is

$$V - [1 - \beta_g(p)](mH + d) - k.$$

The DM anticipates that he will have to pay damages H only if the state is bad, which has probability $[1 - \beta_g(p)]$, and AI supervision reveals that AI is unreliable, which occurs with probability m. It follows that the DM benefits from engaging in AI supervision when

$$H > \frac{k}{(1 - m)[1 - \beta_g(p)]}.$$

Note also that this wasteful AI supervision is more likely to arise when m is lower.

Shared liability. One could also envision a shared liability regime where both the DM and the AI developer are responsible for a portion of the harm, potentially proportionate to their respective percentages of fault. Such a solution would bring in part of the benefits of the two liability regimes we have analyzed. Formally, γ would take a value in $[0, 1]$, with γ now denoting the share of the harm that the DM would be liable for. Efficiency in activity selection and AI supervision improves with higher γ. To see this, consider that all the thresholds identified in Section 3.1 are weakly increasing in γ and Condition (1) is more likely to hold as γ is higher. However, apportioning a higher share of liability to the DM comes at the cost of discouraging the AI developer's investment, which is maximized when $\gamma = 0$. Therefore, the optimal γ depends on the relative weight of ex-ante and ex-post efficiency concerns.

6 Concluding Remarks

Our modeling framework highlights the implications of different regimes that can be envisioned to allocate liability for harm caused by the use of generative AI. If AI is adopted and third parties are harmed, in principle, liability could be allocated to either the developer of the technology or its operator. Liability for AI developers could be made conditional on whether the AI user did not take enough precautions to avoid causing harm.

Ideally, the legal regime should foster the adoption of generative AI when doing so can help improve human decision-making and encourage supervision of the AI system when the technology is not accurate enough. Importantly, as the technology becomes more reliable, the decision maker should save monitoring expenses. Making the AI operator liable would enable the achievement of this first set of objectives. Yet, a trade-off arises when AI accuracy is endogenous. Accountability is what motivates the AI developer to invest in the precision of the technology. We do find that assigning the liability to the AI developer is always desirable, provided that the baseline precision of the technology is sufficiently high. Allocating liability to the AI operator is more likely to be desirable when the human monitoring technology is efficient, when the reputational concerns are minor, and when verifying whether he engaged in AI supervision or the outcome of such activity is difficult. Our findings are policy relevant given the ongoing debate to regulate AI.[29]

We conclude by pointing out some potential limitations of our framework and suggesting some directions for future research. Our analysis has deliberately abstracted from more complete contractual arrangements that might allow for transfers contingent on the occurrence of harm to third parties and the DM's engagement in AI supervision. Although such contractual solutions could theoretically enable the DM and the AI developer to effectively reallocate liability, their feasibility hinges on the legal regimes acting as default rules, which can be contracted around, rather than immutable ones.[30] In the setting we have examined the practical implementation of such stipulations is questionable. The skepticism stems not only from inherent technical limitations, such as the verifiability of AI supervision previously discussed, but also from common explicit prohibitions against the use of contractual provisions to limit an economic operator's exposure to product liability.[31] Notwithstanding these challenges, it is noteworthy that certain AI developers have incorporated disclaimers of responsibility by acknowledging the imperfections of their models and urging users to verify the information provided. These disclaimers may

[29]See, for instance, The Financial Times article *"The global race to set the rules for AI"* published on September 13, 2023.

[30]For a definition of default and immutable rules, see Ayres (1998).

[31]See, for instance, the EU Revised Product Liability Directive, which entered into force on December 9, 2024, and covers AI systems.

serve to waive some degree of legal responsibility in the event of harm. Furthermore, in our model, both the AI developer and the DM can fully compensate the third parties in the case of harm. Should either party be judgment-proof, the optimal allocation of liability would change, potentially creating a need for *ex-ante* regulation. While we have also assumed that there is a single AI developer, future works could explore how liability affects entry and whether competition spurs investment in the precision of the technology.

References

Acemoglu, D. and Lensman, T. (2024). Regulating transformative technologies. *American Economic Review: Insights*, 6(3):359–376.

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

Ayres, I. (1998). Default rules for incomplete contracts. *The New Palgrave Dictionary of Economics and the Law*, 1:585–589.

Buiten, M., De Streel, A., and Peitz, M. (2023). The law and economics of AI liability. *Computer Law & Security Review*, 48:105794.

Buiten, M. C. (2024). Product liability for defective AI. *European Journal of Law and Economics*, 57(1):239–273.

Chen, Y. and Hua, X. (2012). Ex ante investment, ex post remedies, and product liability. *International Economic Review*, 53(3):845–866.

Chen, Y. and Hua, X. (2024a). Multimarket firms and product liability: uniform versus variable rules. *The Journal of Law, Economics, and Organization*, page ewae022.

Chen, Y. and Hua, X. (2024b). Product safety in the age of AI: Autonomy, R&D, and ai liability. *HKUST Business School Research Paper*, (2024-158).

Dahl, M., Magesh, V., Suzgun, M., and Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

Dai, T. and Singh, S. (2025). Artificial intelligence on call: The physician's decision of whether to use AI in clinical practice. *Journal of Marketing Research*, page forthcoming.

Daughety, A. F. and Reinganum, J. F. (1995). Product safety: liability, r&d, and signaling. *The American Economic Review*, pages 1187–1206.

Daughety, A. F. and Reinganum, J. F. (2008). Communicating quality: a unified model of disclosure and signalling. *The RAND Journal of Economics*, 39(4):973–989.

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

Galasso, A. and Luo, H. (2022). When does product liability risk chill innovation? evidence from medical implants. *American Economic Journal: Economic Policy*, 14(2):366–401.

Gans, J. S. (2024a). Copyright policy options for generative artificial intelligence. Technical report, National Bureau of Economic Research.

Gans, J. S. (2024b). How will generative AI impact communication? *Economics Letters*, 242:111872.

Gans, J. S. (2025). How learning about harms impacts the optimal rate of artificial intelligence adoption. *Economic Policy*, 40(121):199–219.

Guadalupi, C., Figueroa, N., and Lemus, J. (2024). Regulation and responsible innovation. *Available at SSRN 4984370*.

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

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

Guerreiro, J., Rebelo, S., and Teles, P. (2023). Regulating artificial intelligence. Technical report, National Bureau of Economic Research.

Haupt, C. E. and Marks, M. (2023). AI-generated medical advice—GPT and beyond. *Jama*, 329(16):1349–1350.

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

Henry, E., Loseto, M., and Ottaviani, M. (2022). Regulation with experimentation: Ex ante approval, ex post withdrawal, and liability. *Management Science*, 68(7):5330–5347.

Hua, X. and Spier, K. E. (2020). Product safety, contracts, and liability. *The RAND Journal of Economics*, 51(1):233–259.

Liao, W., Lu, X., Fei, Y., Gu, Y., and Huang, Y. (2024). Generative AI design for building structures. *Automation in Construction*, 157:105187.

Llanes, G. and Madio, L. (2024). Business strategy and regulation of generative AI firms. *Available at SSRN 4933790*.

Shavell, S. (1992). Liability and the incentive to obtain information about risk. *The Journal of Legal Studies*, 21(2):259–270.

Yang, S. A. and Zhang, A. H. (2024). Generative AI and copyright: A dynamic perspective. *arXiv preprint arXiv:2402.17801*.

Appendix A

Tree Diagram and Information Sets

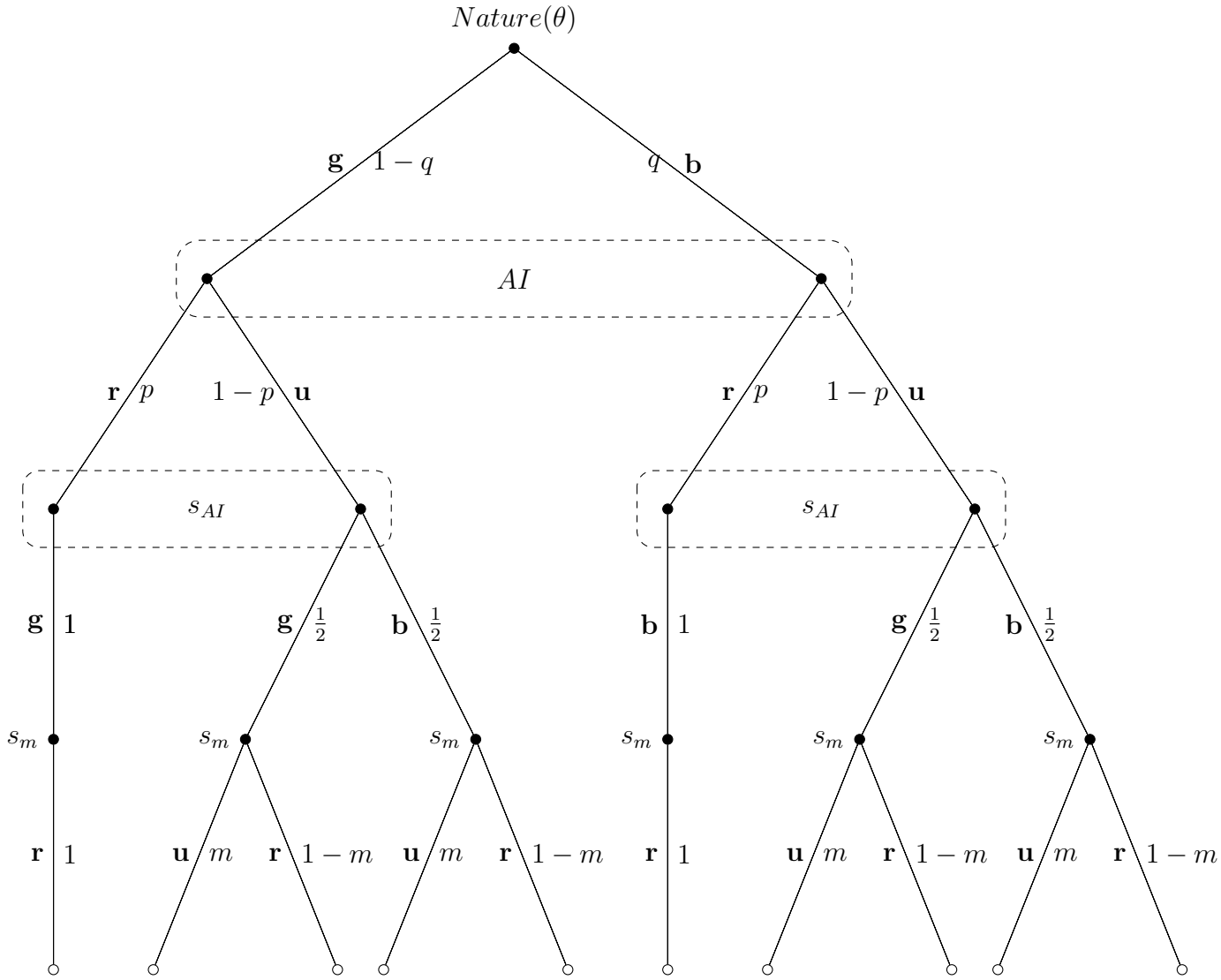

Figure 7: Tree diagram and information sets

Proof of Proposition 1

To prove (a), it suffices to note $\underline{p}_h(0) < \underline{p}_h(1)$.

To prove (b), suppose Condition (1) is not satisfied for $\gamma = 0$. See that $p_A(0) > \underline{p}_h(1)$ if and only if $m > \hat{m} \equiv \frac{q(kd+HV)}{V[q(H+d)-V]}$. For $\hat{m}$ to be consistent with the assumptions that Condi-

tion (1) is satisfied for $\gamma = 1$ but not for $\gamma = 0$, it must be that $\hat{m} \in \left(\frac{kq(d+H)}{V[q(H+d)-V]}, \frac{kqd}{V[qd-V]}\right)$. This occurs as long as $q \in \left(\frac{V}{d}, \frac{V^2}{d(V-k)}\right]$ and $V > k$.

□

Proof of Remark 3

For any expectation $p^e \in [p_0, 1]$ that the DM may hold over the accuracy of the AI technology, the maximum price the DM is willing to pay at stage 2 to purchase the AI system is $\pi_h^{DM}(p^e, T(p^e)) = 0$. Then, for any $T(p^e)$, the AI developer will choose $p^{DM} = \arg\max_{p \in [p_0,1]} T(p^e) - c(p - p_0, p_0) = p_0$, because the actual choice of p is privately observed by the AI developer and cannot influence the DM's expectation. As the DM forms his expectation rationally, this requires that in equilibrium the DM will correctly predict $p^e = p^{DM}$. □

Proof of Lemma 1

The AI developer chooses (T, p) to maximize Π subject to the DM's acceptance.

First, suppose $q > V/d$ and Condition (1) is satisfied for $\gamma = 0$. If $p^{nm} > \overline{p}_h(0)$ and $p^m \notin [\underline{p}_h(0), \overline{p}_h(0)]$, the AI developer sets T such that the DM's participation constraint binds and chooses p to maximize $T(p^e) - c(p - p_0, p_0) - \frac{q(1-p)H}{2}$. This yields $p^{AI} = p^{nm}$ if $T(p^{nm}) - c(p^{nm} - p_0, p_0) - \frac{q(1-p^{nm})H}{2} \geq 0$ and $p^{AI} = p_0$ otherwise, where

$$T(p^{nm}) = \frac{(1-q)(1+p^{nm})}{2}V + \frac{q(1-p^{nm})}{2}(d-V).$$

Also note that if $T(p^{nm}) - c(p^{nm} - p_0, p_0) - \frac{q(1-p^{nm})H}{2} < 0$, then $T(p_0) - \frac{q(1-p^{nm})H}{2} < 0$, where $T(p_0) = \frac{(1-q)(1+p_0)}{2}V + \frac{q(1-p_0)}{2}(d-V)$ because $T(p^{nm}) > T(p_0)$ and $p^{nm} \equiv \arg\min_p c(p - p_0, p_0) + \frac{q(1-p)H}{2}$.

If $p^{nm} > \overline{p}_h(0)$ and $p^m \in [\underline{p}_h(0), \overline{p}_h(0)]$, the AI developer sets T such that the DM's participation constraint binds, and the developer compares the profit she can get if she induces AI supervision or not. In the former case, she chooses p to maximize $T(p^e) - c(p - p_0, p_0) - \frac{q(1-p)(1-m)H}{2}$. The solution to this maximization problem is $p = p^m$. In the latter case, she chooses p to maximize $T(p^e) - c(p - p_0, p_0) - \frac{q(1-p)H}{2}$. The solution to this maximization problem yields $p = p^{nm}$. The AI developer weakly prefers p^{nm} to p^m if:

$$\frac{qH}{2}[(1-p^m)(1-m) - (1-p^{nm})] \geq c(p^{nm} - p_0, p_0) - c(p^m - p_0, p_0).$$

If the above inequality is satisfied, $p^{AI} = p^{nm}$ if $T(p^{nm}) - c(p^{nm} - p_0, p_0) - \frac{q(1-p^{nm})H}{2} \geq 0$, and $p^{AI} = 0$ otherwise. If the above inequality is not satisfied, $p^{AI} = p^m$ if $T(p^m) - c(p^m - p_0, p_0) - \frac{q(1-p^m)(1-m)H}{2} \geq 0$, where

$$T(p^m) = (1-q)\left[p^m + \frac{1-p^m}{2}(1-m)\right]V - \frac{q(1-p^m)(1-m)}{2}(d-V) - \frac{(1-q)(1+p^m)+q(1-p^m)}{2}k.$$

Otherwise, $p^{AI} = p^{nm}$ if $T(p^{nm}) - c(p^{nm} - p_0, p_0) - \frac{q(1-p^{nm})H}{2} \geq 0$ or else $p^{AI} = p_0$.

If $p^{nm} \leq \overline{p}_h(0)$ and $p^m \in [\underline{p}_h(0), \overline{p}_h(0)]$, then following the similar reasoning as in the prior case, $p^{AI} = p^m$ if $T(p^m) - c(p^m - p_0, p_0) - \frac{q(1-p^m)(1-m)H}{2} \geq 0$, and $p^{AI} = p_0$ otherwise. If $p^{nm} \leq \overline{p}_h(0)$ and $p^m < \underline{p}_h(0)$, then $p^{AI} = p_0$.

Next suppose $q > V/d$ and Condition (1) is not satisfied for $\gamma = 0$. If $p^{nm} > p_A(0)$, the AI developer sets T such that the DM's participation constraint binds and chooses p to maximize $T(p^e) - c(p - p_0, p_0) - \frac{q(1-p)H}{2}$. This yields $p^{AI} = p^{nm}$ if $T(p^{nm}) - c(p^{nm} - p_0, p_0) - \frac{q(1-p^{nm})H}{2} \geq 0$ and $p^{AI} = p_0$ otherwise. If $p^{nm} \leq p_A(0)$, then $p^{AI} = p_0$.

Finally, suppose $q \in [\hat{q}, V/d]$. In this case, the AI developer sets T such that the DM's participation constraint binds and chooses p to maximize $T(p^e) - c(p - p_0, p_0) - \frac{q(1-p)H}{2}$. This yields $p^{AI} = p^{nm}$ if $T(p^{nm}) - c(p^{nm} - p_0, p_0) - \frac{q(1-p^{nm})H}{2} \geq 0$, and $p^{AI} = p_0$ otherwise. □

Proof of Proposition 2

First, we identify the conditions under which AI-Developer liability can strictly dominate AI-Operator liability. This requires that $p^{AI} > p_0$. When $q \in [\hat{q}, V/d]$, this happens if

$$T(p^{nm}) - c(p^{nm} - p_0, p_0) - \frac{q(1-p^{nm})H}{2} \geq 0. \tag{A1}$$

If $q > V/d$ and Condition (1) is not satisfied for $\gamma = 0$, (A1) must hold and $p^{nm} > p_A(0)$.

If $q > V/d$ and Condition (1) is satisfied for $\gamma = 0$, either (A1) holds and $p^{nm} > \overline{p}_h(0)$ or

$$T(p^m) - c(p^m - p_0, p_0) - \frac{q(1-p^m)(1-m)H}{2} \geq 0$$

and $p^m \in [\underline{p}_h(0), \overline{p}_h(0)]$.

Now, we proceed by proving a series of claims.

Claim 1. *If Condition (1) is not satisfied for $\gamma = 1$, AI-Developer liability always leads to weakly higher profits than AI-Operator liability.*

Proof. Note that by assuming that the AI developer makes a take-it-or-leave-it offer, industry profits coincide with the AI developer's profits. This is true under both regimes. If Condition (1) does not hold for $\gamma = 1$, expected industry profits under AI-Operator liability, denoted Π^{Op}, are

$$\Pi^{Op} = \max\left\{\frac{(1-q)(1+p_0)}{2}V - \frac{q(1-p_0)}{2}(d+H-V), 0\right\}. \tag{A2}$$

Expected industry profits under AI-Developer liability, denoted Π^{Dev}, are

$$\Pi^{Dev} = \max\left\{\frac{(1-q)(1+p^{AI})}{2}V - \frac{q(1-p^{AI})}{2}(d+H-V) - c(p^{AI}-p_0, p_0), 0\right\}. \tag{A3}$$

If $p^{AI} = p_0$, the two expressions coincide. Now, define

$$\Delta p^*_{nm}(p_0) := \arg\max_{\Delta p \in [0, 1-p_0]} \quad \frac{(1-q)(1+\Delta p+p_0)}{2}V - \frac{q(1-\Delta p-p_0)}{2}(d+H-V) - c(\Delta p, p_0);$$

that is, $\Delta p^*_{nm}(p_0)$ is the unique level of Δp that, for any given p_0, maximizes the surplus of using the AI system when AI monitoring does not take place. Because the maximand is increasing in Δp in the interval $(0, \Delta p^*_{nm}(p_0))$ and $\Delta p^{nm} \in (0, \Delta p^*_{nm}(p_0))$, it must be that $\Pi^{Dev} > \Pi^{Op}$ whenever $p^{AI} = p^{nm}$. □

Claim 2. *If Condition (1) is satisfied for $\gamma = 1$ and $q \in [\hat{q}, V/d]$, AI-Operator liability may lead to strictly higher profits than AI-Developer liability only if $p_0 \in [\underline{p}_h(1), \overline{p}_h(1)]$.*

Proof. If $q \in [\hat{q}, V/d]$, Π^{Dev} are exactly as in (A3). If Condition (1) is satisfied for $\gamma = 1$ and $p_0 \notin [\underline{p}_h(1), \overline{p}_h(1)]$, Π^{Op} are as in (A2). As shown in the previous claim, industry profits are always weakly higher under AI-Developer liability. If Condition (1) is satisfied for $\gamma = 1$ and $p_0 \in [\underline{p}_h(1), \overline{p}_h(1)]$, Π^{Op} are given by:

$$\Pi^{Op} = (1-q)\left[p_0 + \frac{1-p_0}{2}(1-m)\right]V - \frac{q(1-p_0)(1-m)}{2}(d+H-V) - \frac{(1-q)(1+p_0)+q(1-p_0)}{2}k, \tag{A4}$$

which is greater than (A3) if m/k is sufficiently large relative to Δp^{AI}. □

Claim 3. *If Condition (1) is satisfied for $\gamma = 1$ but not for $\gamma = 0$, AI-Operator liability may lead to strictly higher profits than AI-Developer liability only if $p_0 \in [\underline{p}_h(1), \overline{p}_h(1)]$.*

Proof. If Condition (1) is not satisfied for $\gamma = 0$, for any $q \geq \hat{q}$, Π^{Op} are as in (A2). Hence, the same result shown in the previous claim holds. □

Claim 4. *If Condition (1) is satisfied for $\gamma = 0$ and $q > V/d$, AI-Operator liability may lead to strictly higher profits than AI-Developer liability only if $p_0 \in (\overline{p}_h(0) - \Delta p^m, \overline{p}_h(1)]$.*

Proof. Recall that $\underline{p}_h(0) < \underline{p}_h(1)$ and when AI monitoring takes place under AI-Developer liability, expected industry profits are

$$\begin{aligned}\Pi^{Dev} =&(1-q)\left[p^m + \frac{1-p^m}{2}(1-m)\right]V - \frac{q(1-p^m)(1-m)}{2}(d+H-V)\\ &-\frac{(1-q)(1+p^m)+q(1-p^m)}{2}k - c(p^m - p_0, p_0).\end{aligned} \tag{A5}$$

Define

$$\begin{aligned}\Delta p_m^*(p_0) := \arg\max_{\Delta p \in [0,1-p_0]} \quad &(1-q)\left[p_0 + \Delta p + \frac{1-p_0-\Delta p}{2}(1-m)\right]V\\ &-\frac{q(1-p_0-\Delta p)(1-m)}{2}(d+H-V)\\ &-\frac{(1-q)(1+p_0+\Delta p)+q(1-p_0-\Delta p)}{2}k - c(\Delta p, p_0);\end{aligned}$$

that is, $\Delta p_m^*(p_0)$ is the unique level of Δp that, for any given p_0, maximizes the surplus of using the AI system when AI monitoring takes place. As the maximand is increasing in Δp in the interval $(0, \Delta p_m^*(p_0))$ and $\Delta p^m \in (0, \Delta p_m^*(p_0))$, it must be that $\Pi^{Dev} > \Pi^{Op}$ whenever monitoring takes place under both liability regimes. Recall that AI supervision takes place under AI-Developer liability as long as $p^m = p_0 + \Delta p^m \leq \overline{p}_h(0)$. For levels of p_0 above this threshold and below or equal to $\overline{p}_h(1)$, the comparison is between (A3) and (A4), and as stated in Claim 2, the latter is greater when m/k is sufficiently large relative to Δp^{AI}. For values of p_0 above $\overline{p}_h(1)$, we have found in Claim 1 that (A3) is weakly higher than (A2). □

Claim 5. *There exists $\tilde{p}_0 < \overline{p}_h(1)$ such that for any $p_0 > \tilde{p}_0$, AI-Developer liability leads to strictly higher profits.*

Proof. Note that Π^{Op} is a continuous and weakly increasing function of p_0 in the interval $[0,1]$ and Π^{Dev} is a continuous and weakly increasing function of p_0 in the interval $(\overline{p}_h(0) - \Delta p^m, 1]$. In the previous claims, we have shown that AI-Operator liability may lead to strictly higher profits only within the interval $[\underline{p}_h(1), \overline{p}_h(1)]$. In fact, when $p_0 \uparrow \overline{p}_h(1)$, AI-Developer liability generates strictly higher profits than AI-Operator liability because (A3) is positive and strictly higher than (A4). To see this, consider that (A4) is weakly lower than (A2) evaluated at $p_0 \downarrow \overline{p}_h(1)$ and (A3) is higher than (A2). As $\overline{p}_h(1) < 1$, the claim follows. □

Proof of Proposition 3

Start with the effect of an increase in d. Suppose (1) is not satisfied for $\gamma = 0$ and d is initially such that $q \in [\hat{q}, V/d]$. Let (A1) hold. It follows that $p^{AI} = p^{nm}$. Now suppose d increases. As $T(p^{nm})$ increases in d, (A1) continues to hold. However, if $p^{nm} < p_A(0)$, the AI developer will choose $p^{AI} = p_0$.

As for the effect of an increase in m on the choice of p, suppose $q > V/d$ and Condition (1) is satisfied for $\gamma = 0$, and $p^{nm} > \overline{p}_h(0)$, whereas $p^m \in [\underline{p}_h(0), \overline{p}_h(0)]$. The AI developer weakly prefers p^{nm} to p^m if

$$\frac{qH}{2}[(1-p^m)(1-m) - (1-p^{nm})] \geq c(p^{nm} - p_0) - c(p^m - p_0).$$

Note, a marginal increase in m makes it more difficult to satisfy the above inequality, which is reported in the text as (5). To see this, let

$$Z := \frac{qH}{2}[(1-p^m)(1-m) - (1-p^{nm})] - \Big[c(p^{nm} - p_0) - c(p^m - p_0)\Big].$$

It is easy to see that

$$\frac{\partial Z}{\partial m} = -\frac{qH}{2}(1-p^m) - \frac{\partial p^m}{\partial m} \underbrace{\left[\frac{qH}{2}(1-m) - c_p(p^m - p_0)\right]}_{=0},$$

where the latter term is equal to 0 because of the envelope theorem. The implication is that an increase in m may increase the likelihood of the AI developer opting for an inefficiently lower level of AI system accuracy as (5) is not satisfied, and hence, the AI developer chooses $p = p^m$ but it holds that

$$T(p^{nm}) - c(p^{nm} - p_0) - \frac{q(1-p^{nm})H}{2} > T(p^m) - c(p^m - p_0) - \frac{q(1-p^m)(1-m)H}{2}.$$

We conclude with the effect of p_0 on industry profits. Consider the case in which Condition (1) is satisfied for $\gamma = 0$ and $q > V/d$. Suppose p_0 slightly increases from p_0' to p_0''. Suppose further that $p_0' \in (\underline{p}_h(0) - \Delta p^m, \overline{p}_h(0) - \Delta p^m)$, whereas $p_0'' \in (\overline{p}_h(0) - \Delta p^m, \overline{p}_h(1))$. The increase in p_0 may reduce industry profits because at p_0' (A5) can be greater than (A3) at p_0''. Moreover, (A4) can be greater than (A3) at p_0'', implying that a liability regime change is desirable. Yet, it may be that (A5) at p_0' is still greater than (A4) at p_0''. □

Appendix B

Additional Comparative Statics on Price and Adoption of Generative AI

The purpose of this subsection is to perform some comparative statics to analyze the impact of some crucial parameters on the price and adoption of the Generative AI system.

Remark 6. *Under both liability regimes,*

(a) an increase in the baseline accuracy of the technology, p_0, favors AI adoption and leads to higher prices.

(b) an increase in the efficiency of monitoring, m/k, fosters AI adoption, widens the scope of AI supervision, and leads to higher prices.

(c) an increase in reputational concerns, d, discourages adoption of the AI system, and reduces prices.

An increase in H always discourages adoption of the AI system and reduces prices under AI-Operator liability, but may lead to higher adoption and higher prices under AI-Developer liability.

Proof. We begin by considering AI-Operator liability. Suppose AI supervision is not efficient, that is, Condition (1) does not hold for $\gamma = 1$. From (3) and knowing that $p^{DM} = p_0$, we can retrieve the price charged by the AI developer:

$$T^{nm} = \begin{cases} 0, & \text{if } p_0 \leq p_A(1); \\ \frac{(1-q)(1+p_0)}{2}V - \frac{q(1-p_0)}{2}(d+H-V), & \text{if } p_0 > p_A(1). \end{cases}$$

We can see that an increase in d or H reduces T^{nm} and hinders adoption of the AI system, where the latter owes to increasing the threshold $p_A(1)$, above which the DM is willing to use the technology; conversely, an improvement in the baseline technology p_0 boosts price and adoption.

Suppose now that AI supervision is efficient, that is, Condition (1) holds for $\gamma = 1$. From (2), and knowing that $p^{DM} = p_0$, we can retrieve the price charged by the AI

developer: $T^m =$

$$\begin{cases} 0, & \text{if } p_0 < \underline{p}_h(1); \\ (1-q)\left[p_0 + \frac{1-p_0}{2}(1-m)\right]V - \frac{q(1-p_0)(1-m)}{2}(d+H-V) - \frac{(1-q)(1+p_0)+q(1-p_0)}{2}k, & \text{if } p_0 \in \left[\underline{p}_h(1), \overline{p}_h(1)\right] \\ \frac{(1-q)(1+p_0)}{2}V - \frac{q(1-p_0)}{2}(d+H-V), & \text{if } p_0 > \overline{p}_h(1). \end{cases}$$

An increase in d or H reduces T^m and hinders adoption of the AI system, where the latter owes to rising the threshold $\underline{p}_h$; an increase in the efficiency of monitoring, m/k, leads to higher prices, fosters adoption by reducing the threshold $\underline{p}_h(1)$ and widens the scope of AI supervision by increasing $\overline{p}_h(1)$. The effect of an increase in p_0 on transfers is always positive. This is immediate to see when $p > \overline{p}_h(1)$, whereas is less apparent when $p_0 \in \left[\underline{p}_h(1), \overline{p}_h(1)\right]$. In particular, in this range

$$\frac{\partial T^m}{\partial p_0} = (1-q)\left(1 - \frac{1-m}{2}\right)V + \frac{q(1-m)}{2}(d+H-V) - \frac{(1-2q)}{2}k.$$

However, if it were negative, T^m would also be negative.

See also that Condition (1) is easier to satisfy when m/k, d, or H take higher values.[32]

Next, we consider AI-Developer liability. Suppose Condition (1) does not hold for $\gamma = 0$. From (2), and knowing that $p^{AI} \in \{p_0, p^{nm}\}$, we can retrieve the price charged by the AI developer:

$$T_{AI}^{nm} = \begin{cases} 0, & \text{if } p^{AI} \leq p_A(0). \\ \frac{(1-q)(1+p^{AI})}{2}V - \frac{q(1-p^{AI})}{2}(d-V), & \text{if } p^{AI} > p_A(0). \end{cases}$$

An increase in d reduces T_{AI}^{nm} and hinders adoption of the AI system, where the latter owes to rising the threshold $p_A(0)$, above which the DM is willing to use the technology; conversely, an improvement in the baseline technology p_0 boosts price and adoption. Note an increase in H does not affect the threshold $p_A(0)$ and has a positive impact on price when $p^{AI} = p^{nm}$ by increasing AI precision. Its effect on adoption is however ambiguous, because it makes it harder to satisfy the participation constraint.

[32]Note that this does not change the direction of the effects of these parameters on transfers because when they entail a shift from T^{nm} to T^m, they simultaneously imply a change in the sign of the comparison between $p_A(1)$ and $\overline{p}_h(1)$.

Suppose instead that Condition (1) holds for $\gamma = 0$. From (3), and knowing that $p^{AI} \in \{p^m, p^{nm}\}$, we can retrieve the price charged by the AI developer: $T^m_{AI} =$

$$\begin{cases} 0, & \text{if } p^{AI} < \underline{p}_h(0). \\ (1-q)\left[p^m + \frac{1-p^m}{2}(1-m)\right]V - \frac{q(1-p^m)(1-m)}{2}(d-V) - \frac{(1-q)(1+p^m)+q(1-p^m)}{2}k, & \text{if } p^{AI} \in \left[\underline{p}_h(0), \overline{p}_h(0)\right]; \\ \frac{(1-q)(1+p^{nm})}{2}V - \frac{q(1-p^{nm})}{2}(d-V), & \text{if } p^{AI} > \overline{p}_h(0). \end{cases}$$

We can see an increase in d reduces T^m_{AI} and hinders adoption of the AI system, as it rises the threshold $\underline{p}_h(0)$; an increase in the efficiency of monitoring, m/k, leads to higher prices, fosters adoption by reducing the threshold $\underline{p}_h(0)$, and widens the scope of AI supervision by increasing $\overline{p}_h(0)$. The effect of an increase in p_0 and/or H on transfers is always positive. Higher baseline precision helps adoption, whereas a larger H has an ambiguous effect: on the one hand, it makes it harder to satisfy the participation constraint; on the other hand, it increases the surplus by boosting AI precision. □

To provide an intuition, we focus on the case of AI-Operator liability. For the most part, the same logic applies to the case of AI-Developer liability and we highlight the only qualitative difference at the end of this paragraph. An increase in the baseline accuracy of the technology, p_0, increases the likelihood that the DM will find using the AI system is beneficial and increase the price he is willing to pay for its use. An increase in the efficiency of monitoring, m/k, fosters adoption by reducing the threshold $\underline{p}_h(1)$ and widens the scope of AI supervision by increasing $\overline{p}_h(1)$. An increase in reputational concerns, d, hinders adoption of the AI system by increasing the thresholds $p_A(1)$ and $\underline{p}_h(1)$, above which the DM is willing to use the technology, leading to a lower price. The main difference between AI Operator and AI-Developer liability regards the impact of H, which in the latter, does not influence the threshold $\underline{p}_h(0)$ but has a positive impact on the precision of the technology, favoring its adoption, and may lead to a higher price. Although a liable AI developer demanding a higher price when the damages are larger may seem unsurprising, we obtain this result also in a setting where the AI developer can fully extract the trade surplus from the AI operator. Then, the reason we get this result is that a higher H increases the equilibrium precision of the AI technology, thereby amplifying the benefit the DM obtains from using the AI system. On the other hand, if H grows above a certain level, the AI developer will not be willing to provide the technology;

that is, the higher transfer the AI developer could receive would not be enough to cover the higher cost of investing in the AI accuracy and the anticipated liability expenses.

AI-Developer Conditional liability

In this section, we provide the formal analysis for the AI-Developer Conditional liability regime.

AI supervision and activity selection. In stage 3, the DM will choose $a = s$ regardless of the precision of the AI signal if $s_{AI} = b$. If $s_{AI} = g$, the DM may decide to pursue $a = x$. As under AI-Operator liability, the DM strictly prefers $a = x$ to $a = s$ when $p > p_A(1)$. Accordingly, if the DM does not engage in supervision after observing $s_{AI} = g$, his expected utility is $\pi_h^{nm}(g)$, which we report below for convenience:

$$\pi_h^{nm}(g) = \begin{cases} V - [1 - \beta_g(p)](H + d), & \text{if } p > p_A(1); \\ 0, & \text{if } p \leq p_A(1). \end{cases}$$

If the DM engages in AI supervision after observing $s_{AI} = g$, he may or may not find that the AI signal was a hallucination. Suppose the DM learns that he cannot rely on the AI signal, that is, $s_m = u$. If the result of AI supervision is verifiable, the DM will choose $a = s$. If the result is not verifiable, the DM could still shift liability to the AI developer and he will select $a = x$ whenever $q < V/d$.

Suppose the DM learns that the AI signal is reliable, that is, $s_m = r$. Then, he believes $\theta = g$ with probability $\beta_g^m(p)$ and, irrespective of whether the signal is verifiable, he selects $a = x$ if $V - [1 - \beta_g^m(p)]d \geq 0$, that is, if $p \geq p_B(0)$. Hence, if $qd < V$, the DM will always choose $a = x$ if AI supervision reveals that the AI system is reliable.

To determine the DM's incentive to engage in AI supervision, we first need to compute his expected utility from double-checking $s_{AI} = g$. We distinguish between two cases.

Case 1. When the result of AI supervision is verifiable or when it is not, but $q > \frac{V}{d}$, the DM's expected utility from supervising AI is:

$$\pi_h^{mAI}(g) = -k + \begin{cases} \beta_g(p)\left(1 - \frac{m(1-p)}{1+p}\right)V - [1 - \beta_g(p)](1 - m)(d - V), & \text{if } p \geq p_B(0); \\ 0, & \text{if } p < p_B(0). \end{cases}$$

We need to entertain two possibilities. First, if $p \in [p_B(0), p_A(1)]$, the relevant comparison is between:

$$\beta_g(p)\left(1-\frac{m(1-p)}{1+p}\right)V-[1-\beta_g(p)](1-m)(d-V)-k$$

and 0 because the DM selects $a=x$ only if AI supervision does not reveal that $s_{AI}=g$ is a hallucination. As under AI-Developer liability, the DM weakly prefers to engage in AI supervision if $p \geq \underline{p}_h(0)$, which is always higher than $p_B(0)$ and is lower than $p_A(1)$ if $q > \frac{V}{d}$ and Condition (6) is satisfied.

The second possibility arises when $p \in [p_A(1), 1]$. The relevant comparison is between:

$$\beta_g(p)\left(1-\frac{m(1-p)}{1+p}\right)V-[1-\beta_g(p)](1-m)(d-V)-k$$

and $V-[1-\beta_g(p)](d+H)$ because the DM selects $a=x$ when he does not engage in AI supervision. The DM weakly prefers to supervise the AI if:

$$p \leq \overline{p}_h^{AIC}.$$

It is easy to see that $\overline{p}_h^{AIC} < 1$, whereas $\overline{p}_h^{AIC} > p_A(1)$ if Condition (6) is satisfied.

We can now write the DM's expected utility when $s_{AI}=g$. If AI monitoring is inefficient, that is, if Condition (6) is not satisfied, the DM never engages in AI supervision after a favorable AI signal, and his expected utility is $\pi_h^{nm}(g)$. If Condition (6) is satisfied, the DM's expected utility after observing $s_{AI}=g$ is

$$\pi_h^{AIC}(g)=\begin{cases}0, & \text{if } p<\underline{p}_h(0);\\ \beta_g(p)\left(1-\frac{m(1-p)}{1+p}\right)V-[1-\beta_g(p)](1-m)(d-V)-k, & \text{if } p\in[\underline{p}_h(0),\overline{p}_h^{AIC}];\\ V-[1-\beta_g(p)](d+H), & \text{if } p>\overline{p}_h^{AIC}.\end{cases}$$

Case 2. Suppose now the result of AI supervision is not verifiable and $q \in [\hat{q}, V/d]$. The DM's expected utility from supervising AI is

$$\pi_h^{mAI}(g) = -k+\beta_g(p)V+[1-\beta_g(p)](V-d).$$

We compare the previous expression for the DM's expected utility with the one obtained when he does not engage in supervision, $\pi_h^{nm}(g)$. We find two new threshold values that we denote by

$$\underline{p}_{hL}^{AIC} := \frac{V-qd-k}{V-qd-k-2(1-q)(V-k)}, \quad \overline{p}_{hL}^{AIC} := \frac{-qH+k}{-qH+k-2(1-q)k},$$

We find $\underline{p}_{hL}^{AIC} < \underline{p}_h(1)$ and $\overline{p}_{hL}^{AIC} > \overline{p}_h(1)$.

We now analyze the DM's decision to adopt AI for high-risk activities. To this end, we need to specify his expected utility from using the AI system. Let us suppose Condition (6) holds and $q > \frac{V}{d}$. Then, we can distinguish between three parameter regions:

$$\pi_h^{AIC} = -T \tag{B1}$$
$$+ \begin{cases} 0, & \text{if } p < \underline{p}_h(0); \\ (1-q)\left[p + \frac{1-p}{2}(1-m)\right]V - \frac{q(1-p)(1-m)}{2}(d-V) - \frac{(1-q)(1+p)+q(1-p)}{2}k, & \text{if } p \in \left[\underline{p}_h(0), \overline{p}_h^{AIC}\right]; \\ \frac{(1-q)(1+p)}{2}V - \frac{q(1-p)}{2}(d+H-V), & \text{if } p > \overline{p}_h^{AIC}. \end{cases}$$

When minor reputational concerns exist and the result of supervision is verifiable, the expression is the same.

If Condition (6) is not satisfied, the DM's expected utility is given by (3), which we report below for convenience:

$$\pi_h^{AIC} = -T + \begin{cases} 0, & \text{if } p \leq p_A(1). \\ \frac{(1-q)(1+p)}{2}V - \frac{q(1-p)}{2}(d+H-V), & \text{if } p > p_A(1). \end{cases}$$

Lastly, if the result of AI supervision is not verifiable and $q \in \left[\hat{q}, \frac{V}{d}\right]$, the DM's expected utility is

$$\pi_{hL}^{AIC} = -T \tag{B2}$$
$$+ \begin{cases} 0, & \text{if } p < \underline{p}_{hL}^{AIC}; \\ (1-q)\left[p + \frac{1-p}{2}\right]V + \frac{q(1-p)}{2}(V-d) - \frac{(1-q)(1+p)+q(1-p)}{2}k, & \text{if } p \in \left[\underline{p}_{hL}^{AIC}, \overline{p}_{hL}^{AIC}\right]; \\ \frac{(1-q)(1+p)}{2}V - \frac{q(1-p)}{2}(d+H-V), & \text{if } p > \overline{p}_{hL}^{AIC}. \end{cases}$$

Proof of Remark 4

It follows directly from the comparisons in the main text. □

Proof of Proposition 4

To identify the conditions under which AI-Operator liability can maximize profits, we first note they will be at least as stringent as in Proposition 2. We now show that if the

result of AI supervision is verifiable or it is not but $q > V/d$, AI-Operator liability results in lower industry profits than either AI-Developer liability or AI-Developer Conditional liability. We begin by proving the following lemma.

Lemma 2. *If p is privately observable and engaging in AI supervision is verifiable and either its result is verifiable or $q > V/d$, under AI-Developer Conditional liability, $p^{AIC} \in \{p_0, p^m\}$.*

Proof. The AI developer chooses (T, p) to maximize Π subject to the DM's acceptance. Suppose Condition (6) holds, $p^m \in [\underline{p}_h(0), \overline{p}_h^{AIC}]$, and $T(p^m) - c(p^m, p_0) - \frac{q(1-p^m)(1-m)H}{2} \geq 0$. Then, $p^{AIC} = p^m$. Otherwise, $p^{AIC} = p_0$ because the AI developer would not face liability. □

We now prove the following claim.

Claim 6. *If engaging in AI supervision is verifiable and either its result is verifiable or $q > V/d$, either AI-Developer liability or AI-Developer Conditional liability result in weakly higher profits than AI-Operator liability.*

Proof. First recall that if Condition (1) is satisfied for $\gamma = 1$, so is Condition (6). Because of Claims 3 and 4 of Proposition 2, AI-Operator liability can dominate AI-Developer liability only within the interval $p_0 \in [\underline{p}_h(1), \overline{p}_h(1)]$. Yet, in this interval, industry profits under AI-Operator liability, given by expression (A4), are weakly lower than industry profits under AI-Developer Conditional liability, which are given by

$$\begin{aligned}\Pi^{AIC} =& (1-q)\left[p^{AIC} + \frac{1-p^{AIC}}{2}(1-m)\right]V - \frac{q(1-p^{AIC})(1-m)}{2}(d+H-V) \\ &- \frac{(1-q)(1+p^{AIC}) + q(1-p_0)}{2}k - c(p^{AIC} - p_0, p_0),\end{aligned} \tag{B3}$$

because $p^{AIC} \in \{p_0, p^m\}$ and both $\underline{p}_h(0) < \underline{p}_h(1)$ and $\overline{p}_h^{AIC} > \overline{p}_h(1)$. □

We now consider the case in which the result of AI supervision is not verifiable and $q \in [\hat{q}, V/d]$. First, we determine which levels of investments are feasible.

Lemma 3. *If p is privately observable and engaging in AI supervision is verifiable but its result is not and $q \in [\hat{q}, V/d]$, under AI-Developer Conditional liability, $p_{hL}^{AIC} \in \{p_0, p^{nm}\}$.*

Proof. Note the AI developer's expected liability bill is 0 if $p \neq [\underline{p}_{hL}^{AIC}, \overline{p}_{hL}^{AIC}]$, and $\frac{q(1-p)}{2}H$ otherwise. Thus, when $p^{nm} \in [\underline{p}_{hL}^{AIC}, \overline{p}_{hL}^{AIC}]$ and $T_{hL}(p^{nm}) - c(p^{mn}, p_0) - \frac{q(1-p)H}{2} \geq 0$ that will be the equilibrium precision of the AI system, where

$$T_{hL}(p^{nm}) = (1-q)\left[p + \frac{1-p}{2}\right]V + \frac{q(1-p)}{2}(V-d) - \frac{(1-q)(1+p)+q(1-p)}{2}k.$$

Otherwise, $p_{hL}^{AIC} = p_0$. □

We now provide the following claim:

Claim 7. *If Condition (1) is satisfied for $\gamma = 1$, $q \in [\hat{q}, V/d]$, engaging in AI supervision is verifiable but its result is not, AI-Operator liability may lead to higher profits than AI-Developer liability and AI-Developer Conditional liability only if $p_0 \in [\underline{p}_h(1), \overline{p}_h(1)]$.*

Proof. It immediately follows from Claim 2 of Proposition 2 and noticing that industry profits under AI-Developer Conditional liability would be weakly lower than those under AI-Developer liability, reported in (A3), because of the positive supervision cost only incurred under the AI-Developer Conditional liability regime. □

We conclude by observing that whenever p_0 is such that $p^m > \overline{p}_h^{AIC}$, AI-Developer liability leads to strictly higher profits than the other two regimes. □

Further Extensions

Proof of Remark 5

Let p be observable. Irrespective of the value of the parameters, AI is adopted whenever it is efficient. To see why, let W_h denote the expected welfare that can be achieved from adopting AI for high-risk activities, for any given accuracy of the AI technology. Because of Remark 1, $W_h = \pi_h + T$. Note that adopting AI is efficient whenever $\pi_h + T > 0$ and the AI developer will choose T to satisfy the DM's participation constraint. If $T > 0$, the AI developer will then choose p to maximize $\Pi = T(p) - c(p - p_0, p_0)$, which coincides with the expression that would be maximized by a benevolent social planner. □

No-liability Benchmark

We denote the equilibrium choice under no-liability by the superscript NL, and we prove two claims below.

Claim 8. *The DM's incentives for AI supervision and activity selection are the same under AI-Developer liability and under no liability.*

Proof. Such incentives only depend on the DM's expected payoff. With no liability, the DM would not pay damages H, but would only suffer a reputational damage d if the state turns out to be bad. Hence, his expected payoffs coincide with those reported in Section 3.1 when $\gamma = 0$. □

Claim 9. *Under no liability, the equilibrium accuracy $p^{NL} = p^{DM} = p_0 = p^e$.*

Proof. With no liability, the AI developer need not compensate third parties for the harm they suffer. Hence, she chooses $p^{NL} = \arg\max_{p \in [p_0,1]} T(p^e) - c(p - p_0, p_0) = p_0 = p^{DM}$. □

Harm on DM

Let $\alpha \in [0, 1]$ be the share of harm H suffered by the DM, in addition to the reputational damage d, if activity x is chosen in state $\theta = b$. Third parties continue to suffer $(1 - \alpha)H$ if x is chosen in state $\theta = b$.

We begin by comparing AI-Operator and AI-Developer liability and we prove the following claims.

Claim 10. *The DM's incentives for AI supervision and activity selection under both AI-Operator and AI-Developer liability do not vary with the share $\alpha \in [0, 1]$ of the harm incurred by the DM.*

Proof. Such incentives only depend on the DM's expected payoff. The DM's loss if x is chosen in the bad state is $\alpha H + d$, whereas third parties lose $(1 - \alpha)H$. However, while d cannot be recouped, damages H can be shifted through liability. Thus, the expected disutility to the DM if x is chosen in the bad state is $(\gamma\alpha H + d) + [(1-\gamma)(1-\alpha)H] = \gamma H + d$. Thus, we retrieve the same expected loss as in Section 3.1. □

Claim 11. *The AI developer's investment under both AI-Operator and AI-Developer liability does not vary with the share $\alpha \in [0, 1]$ of the harm incurred by the DM.*

Proof. It is immediate to see that the AI developer maximizes the same expected utility expressions as in Section 4, irrespective of α. □

We now show that the need for a liability regime is greater the larger the share $1-\alpha$ of harm suffered by third parties.

Claim 12. *Under no liability, the DM's incentives for AI supervision and activity selection are closer to social efficiency the larger the share $\alpha \in [0,1]$ of the harm incurred by the DM.*

Proof. Such incentives only depend on the DM's expected payoff and recall from Remark 1 that these are socially efficient if the DM bears $H+d$ if x is chosen in the bad state. Under no liability, the expected loss to the DM is $\alpha H + d$. Thus, if $\alpha = 0$, the DM's incentives for AI supervision and activity selection are the same as under AI-Developer liability. If $\alpha = 1$, they are socially efficient. It is easy to see that the inefficiency is smaller the larger α: take the thresholds $p_A(\gamma), p_B(\gamma), \underline{p}_h(\gamma), \bar{p}_h(\gamma)$, replace γ with α and notice that:

$$\frac{\partial p_A(\alpha)}{\partial \alpha} \geq 0; \quad \frac{\partial p_B(\alpha)}{\partial \alpha} \geq 0; \quad \frac{\partial \bar{p}_h(\alpha)}{\partial \alpha} \geq 0; \quad \frac{\partial \underline{p}_h(\alpha)}{\partial \alpha} \geq 0.$$

Moreover, see that

$$\frac{m}{k} \geq \frac{q(\alpha H + d)}{V[q(\alpha H + d) - V]},$$

which replaces Condition (1) for the case of No liability when the DM bears a fraction α of the harm, is easier to satisfy when α is greater. □

This leads to the following remark.

Remark 7. *If the DM incurs all the harm, i.e., $\alpha = 1$, No liability and AI-Operator liability lead to the same outcome.*

Proof. The above claims have shown that, under AI-Operator liability, the DM's choices are independent of α and that, under no liability, if $\alpha = 1$ the DM's choices are the same as under AI-Operator liability. In both cases, the AI developer would choose $p = p_0$ as she would maximize the same expected utility. Therefore, if the DM suffers all the harm there is no need to make the AI operator liable. □

Low-Risk Activities

Below, we compare AI-Operator and AI-Developer liability when activities are low risk. We use the subscript l to refer to low-risk activities.

AI Supervision and Activity Selection

As stated in Proposition 5, DM's incentives for AI supervision and activity selection are the same in the two regimes.

If in stage 3 $s_{AI} = g$, the DM will choose $a = x$ regardless of the precision of the AI signal. If $s_{AI} = b$, the DM may decide to pursue $a = x$. Before making this choice, the DM may decide to double-check the AI signal. If he does not, he expects to get

$$\max\{V - [1 - \beta_b(p)](H + d), 0\}.$$

The DM strictly prefers $a = x$ to $a = s$ when $p \leq p_{Al}$, where[33]

$$p_{Al} := \frac{V - q(H + d)}{V - q(H + d) + 2q(H + d - V)}.$$

If the DM does not engage in supervision after observing $s_{AI} = b$, his expected utility is

$$\pi_l^{nm}(b) = \begin{cases} V - [1 - \beta_b(p)](H + d), & \text{if } p \leq p_{Al}; \\ 0, & \text{if } p > p_{Al}. \end{cases}$$

If the DM engages in AI supervision after observing $s_{AI} = b$, he may or may not find that the AI signal was a hallucination. If $s_m = u$, the DM learns that he cannot rely on the AI signal and he will pursue activity $a = x$. If $s_m = r$, the DM believes that $\theta = g$ with probability $\beta_b^m(p)$ and selects $a = x$ if $V - [1 - \beta_b^m(p)](H + d) \geq 0$, that is, if $p \leq p_{Bl}$, where

$$p_{Bl} := \frac{(1 - m)[V - q(H + d)]}{(1 - m)[V - q(H + d)] + 2q(H + d - V)}.$$

It is possible to see that $p_{Bl} = p_{Al}$ when $m = 0$ and p_{Bl} is decreasing in m when $q < \hat{q}$. Intuitively, the DM would always select the risky activity when the AI supervision finds evidence of a hallucination. When AI supervision suggests the AI recommendation is reliable, the DM would be deterred from pursuing the risky activity unless AI accuracy is low. The threshold is decreasing in the precision of human monitoring.

[33] Note the numerator is always non-negative when $q \leq \hat{q}$.

If AI precision is too low, the DM will undertake the risky activity regardless of the outcome of AI supervision. If the AI system and the AI supervision are sufficiently precise, that is, if $p > p_{Bl}$, the DM will undertake the activity only if he finds out it was a hallucination. The DM anticipates that with probability $\beta_b(p)$ the AI system produced $s_{AI} = b$ when $\theta = g$ and AI supervision detects the hallucination with probability m. If so, by undertaking $a = x$, the DM would get V. Moreover, the DM knows that even when $\theta = b$, AI supervision may reveal that the unfavorable signal was the product of an AI hallucination, in which case he would lose $d + H - V$ by choosing $a = x$. This latter instance occurs with probability $Pr[s_m = u|\theta = b \cap s_{AI} = b] = \frac{m(1-p)}{1+p}$. As a result, the DM's expected utility from supervising the AI system when $s_{AI} = b$ is given by

$$\pi_l^m(b) = -k + \begin{cases} V - [1 - \beta_b(p)](H + d), & \text{if } p \leq p_{Bl}; \\ \beta_b(p)mV - [1 - \beta_b(p)]\frac{m(1-p)}{1+p}(H + d - V), & \text{if } p > p_{Bl}. \end{cases}$$

We now analyze the DM's incentive to engage in AI supervision after observing $s_{AI} = b$. If $p \leq p_{Bl}$, the DM never double-checks the AI signal: the DM will choose $a = x$ when $p \in [0, p_{Bl}]$. For $p \in (p_{Bl}, p_{Al}]$, the relevant comparison is between

$$\beta_b(p)mV - [1 - \beta_b(p)]\frac{m(1-p)}{1+p}(H + d - V) - k$$

and $V - [1 - \beta_b(p)](H + d)$. The DM weakly prefers to supervise the AI if

$$p > \underline{p}_l := \frac{(1-m)[V - q(H+d)] + k}{(1-m)[V - q(H+d)] + k + 2q(d + H - V - k)}.$$

One can easily see $\underline{p}_l < 1$ as long as $d + H - V - k > 0$, it is always greater than p_{Bl} as $k > 0$, and is lower than p_{Al} if the following condition is satisfied:

$$\frac{m}{k} \geq \frac{(1-q)(d+H)}{(d + H - V)[V - q(d+H)]}. \tag{B4}$$

For $p > p_{Al}$, the relevant comparison is between

$$\beta_b(p)mV - [1 - \beta_b(p)]\frac{m(1-p)}{1+p}(H + d - V) - k$$

and 0. The DM weakly prefers to supervise the AI if

$$p \leq \overline{p}_l := \frac{m[V - q(H+d)] - k}{m[V - q(H+d)] - k + 2qk}.$$

One can easily see $\overline{p}_l < 1$, whereas $\overline{p}_l > p_{Al}$ if Condition (B4) is satisfied.

We can now write the DM's expected utility when $s_{AI} = b$. If Condition (B4) does not hold, supervision never takes place and the DM's expected utility is $\pi_l^{nm}(b)$. If Condition (B4) holds, his expected utility after observing $s_{AI} = b$ is:

$$\pi_l^{DM}(b) = \begin{cases} V - [1 - \beta_b(p)](d + H), & \text{if } p < \underline{p}_l; \\ \beta_b(p)mV - [1 - \beta_b(p)]\frac{m(1-p)}{1+p}(H + d - V) - k, & \text{if } p \in [\underline{p}_l, \overline{p}_l]; \\ 0, & \text{if } p > \overline{p}_l. \end{cases}$$

Thus, receiving an unfavorable recommendation dissuades the DM from undertaking the risky activity when the AI accuracy is high enough. For intermediate levels of accuracy, the DM will supervise the AI and pursue the risky activity if he finds evidence of a hallucination. If the AI accuracy is sufficiently low, the DM will discard its recommendation.

AI Adoption and Investment Incentives

We now analyze the DM's decision to adopt AI for low-risk activities. The DM's incentive to adopt the AI system is regime-dependent: if the AI recommendation is favorable but an accident occurs, the DM will always suffer a reputation blow but will have to pay damages only under AI-Operator liability.

We begin by reporting the DM's expected utility from adopting the AI system under AI-Operator liability. Suppose Condition (B4) holds. Then, we can distinguish between three parameter regions:

$$\pi_l^{DM} = -T \tag{B5}$$

$$+ \begin{cases} (1-q)V - q(H + d - V), & \text{if } p < \underline{p}_l. \\ (1-q)\left[\frac{1+p}{2} + \frac{(1-p)m}{2}\right]V - \frac{q(1-p)(1+m)}{2}(d + H - V) - \frac{(1-q)(1-p)+q(1+p)}{2}k, & \text{if } p \in \left[\underline{p}_l, \overline{p}_l\right]; \\ \frac{(1-q)(1+p)}{2}V - \frac{q(1-p)}{2}(d + H - V), & \text{if } p > \overline{p}_l. \end{cases}$$

Interestingly, if the AI signal is sufficiently precise, the DM always follows its recommendation, and his expected utility, net of the transfer paid to the AI developer, is the same as when activities are high risk. When the AI precision is low, the signal realization does not affect the DM's choice. Yet, his expected payoff is positive because he will pursue the activity. For intermediate values of AI precision, the DM will engage in AI supervision

following an unfavorable recommendation and choose the safe activity if he does not find a hallucination.

If Condition (B4) does not hold, AI supervision never takes place and the DM's expected utility from using AI is

$$\pi_l^{DM} = -T + \begin{cases} (1-q)V - q(H+d-V), & \text{if } p \leq p_{Al}. \\ \frac{(1-q)(1+p)}{2}V - \frac{q(1-p)}{2}(d+H-V), & \text{if } p > p_{Al}. \end{cases} \tag{B6}$$

Now consider AI-Developer liability. Suppose Condition (B4) holds. Then, we can distinguish between three parameter regions:

$$\pi_l^{AI} = -T \tag{B7}$$

$$+ \begin{cases} (1-q)V - q(d-V) - \frac{q(1+p)}{2}H, & \text{if } p < \underline{p}_l. \\ (1-q)\left[\frac{1+p}{2} + \frac{(1-p)m}{2}\right]V - \frac{q(1-p)(1+m)}{2}(d-V) - \frac{q(1-p)m}{2}H - \frac{(1-q)(1-p)+q(1+p)}{2}k, & \text{if } p \in \left[\underline{p}_l, \overline{p}_l\right]; \\ \frac{(1-q)(1+p)}{2}V - \frac{q(1-p)}{2}(d-V), & \text{if } p > \overline{p}_l. \end{cases}$$

If Condition (B4) does not hold, AI supervision never takes place and the DM's expected utility from using AI is

$$\pi_l^{AI} = -T + \begin{cases} (1-q)V - q(d-V) - \frac{q(1+p)}{2}H, & \text{if } p \leq p_{Al}. \\ \frac{(1-q)(1+p)}{2}V - \frac{q(1-p)}{2}(d-V), & \text{if } p > p_{Al}. \end{cases} \tag{B8}$$

If p is privately observable, the AI developer has no incentive to invest under AI-Operator liability. By contrast, under AI-Developer liability, she has an incentive to increase AI accuracy whenever she expects to sell the AI system to the DM.

Proof of Proposition 5

It follows from the above analysis. □

Shared Liability

Let $\gamma \in [0,1]$ denote the share of harm the DM is liable for. We prove the following claims.

Claim 13. *The DM's incentives for AI supervision and activity selection are closer to social efficiency the higher* γ.

Proof. The proof is similar to that of Claim 12: all thresholds of Section 3.1 are weakly increasing in γ and Condition (1) is easier to satisfy the higher γ. □

Claim 14. *The AI developer's investment is weakly decreasing in γ.*

Proof. To see this, recall that the AI developer's expected utility at stage 2 is given by:

$$\Pi = T - c(\Delta p, p_0) - (1-\gamma)Pr[H]H,$$

and following the same steps as in Lemma 1, there will be three possible equilibrium values of p as function of γ: p_0, $p^m(\gamma) := p_0 + \Delta p^m(\gamma)$, $p^{nm} := p_0 + \Delta p^{nm}(\gamma)$, where $\Delta p^m(\gamma)$ and $\Delta p^{nm}(\gamma)$ are uniquely defined by:

$$\frac{\partial c(\Delta p^m(\gamma), p_0)}{\partial \Delta p} = \left(\frac{(1-\gamma)(1-m)qH}{2}\right); \quad \textit{and} \quad \frac{\partial c(\Delta p^{nm}(\gamma), p_0)}{\partial \Delta p} = \left(\frac{(1-\gamma)qH}{2}\right).$$

It is immediate to see that both p^{nm} and p^m are decreasing in γ. □

In light of the above two claims, a higher γ improves ex-post efficiency but chills ex-ante investment.

Appendix C

More States of the World

In this appendix, we discuss the robustness of the results when we increase the number of states of the world. Suppose that there are $n \geq 2$ states of the world, that is, $\theta \in \Theta := \{1, 2, ..., n\}$. One state is good, namely $\theta = g$ with probability $1 - q$, whereas all the other states are not (i.e., they are equally bad) with complementary probability q. We denote such states as $\theta \neq g$. A reliable AI system truthfully reveals the state of the world, whereas an unreliable AI system produces a random output. In particular, with probability $1/n$, $s_{AI} = g$ when $AI = u$. As a result,

$$\begin{aligned} Pr[s_{AI} = g|\theta = g] = &\quad p + \frac{1-p}{n} = \frac{1 + (n-1)p}{n}, \text{ and} \\ Pr[s_{AI} \neq g|\theta \neq g] = &\quad p + \frac{(1-p)(n-1)}{n} = \frac{(n-1) + p}{n}. \end{aligned}$$

Note $Pr[s_{AI} = g|\theta = g]$ is decreasing in n and is equal to p as n goes to infinity. Intuitively, if there are more states, the AI system will reveal that the state is good only when it is indeed the case. Conversely, $Pr[s_{AI} \neq g|\theta \neq g]$ is increasing in n and goes to 1 as n goes to infinity, because the AI system will be more unlikely to produce $s_{AI} = g$ when $\theta \neq g$: the chance that an unreliable AI system exactly reports $s_{AI} = g$ gets slimmer when more states of the world are possible.

Let us now see how this assumption affects the DM's belief updating. Upon observing $s_{AI} = g$, the DM believes the true state is $\theta = g$ with probability

$$\begin{aligned} \beta_g(p) = Pr[\theta = g|s_{AI} = g] = &\frac{Pr[s_{AI} = g|\theta = g]Pr(\theta = g)}{Pr[s_{AI} = g|\theta = g]Pr(\theta = g) + Pr[s_{AI} = g|\theta \neq g]Pr(\theta \neq g)} \\ = &\frac{[1 + (n-1)p](1-q)}{[1 + (n-1)p](1-q) + (1-p)q} \in [1-q, 1]. \end{aligned}$$

Note $\beta_g(p)$ is increasing in n and $\lim_{n\to\infty} \beta_g(p) = 1$.

Upon observing $s_{AI} \neq g$, the DM believes that the true state is $\theta = g$ with probability

$$\begin{aligned} \beta_{-g}(p) = Pr[\theta = g|s_{AI} \neq g] = &\frac{Pr[s_{AI} \neq g|\theta = g]Pr(\theta = g)}{Pr[s_{AI} \neq g|\theta = g]Pr(\theta = g) + Pr[s_{AI} \neq g|\theta \neq g]Pr(\theta \neq g)} \\ = &\frac{(1-p)(1-q)(n-1)}{(1-p)(1-q)(n-1) + [(n-1) + p]q} \in [0, 1-q]. \end{aligned}$$

Note $\beta_{-g}(p)$ is increasing in n and $\lim_{n\to\infty} \beta_{-g}(p) = \frac{(1-p)(1-q)}{(1-p)(1-q)+q}$.

The DM further updates his belief about the state of the world if he engages in AI supervision. When $s_{AI} = g$ and the DM undertakes monitoring, two possible events may occur. If $s_m = r$, the DM believes that $\theta = g$ with a probability that can be derived by using the chain rule:

$$\begin{aligned}\beta_g^m(p) =& Pr[\theta = g | s_m = r \cap s_{AI} = g] = \frac{Pr[s_m = r|\theta = g \cap s_{AI} = g]Pr[\theta = g|s_{AI} = g]}{Pr[s_m = r|s_{AI} = g]} \\ =& \frac{Pr[s_m = r|\theta = g \cap s_{AI} = g]Pr[\theta = g|s_{AI} = g]}{Pr[s_m = r|\theta = g \cap s_{AI} = g]Pr[\theta = g|s_{AI} = g] + Pr[s_m = r|\theta \neq g \cap s_{AI} = g]Pr[\theta \neq g|s_{AI} = g]} \\ =& \frac{p(1-q)n + (1-m)(1-p)(1-q)}{p(1-q)n + (1-m)(1-p)(1-q) + (1-m)(1-p)q} \in [\beta_g(p), 1].\end{aligned}$$

Note $\beta_g^m(p)$ is increasing in n and $\lim_{n\to\infty} \beta_g^m(p) = 1$.

If the DM engages in AI supervision after observing $s_{AI} \neq g$, and $s_m = r$, the DM believes the true state is $\theta = g$ with probability

$$\begin{aligned}\beta_{-g}^m(p) =& Pr[\theta = g | s_m = r \cap s_{AI} \neq g] = \frac{Pr[s_m = r|\theta = g \cap s_{AI} \neq g]Pr[\theta = g|s_{AI} \neq g]}{Pr[s_m = r|s_{AI} \neq g]} \\ =& \frac{Pr[s_m = r|\theta = g \cap s_{AI} \neq g]Pr[\theta = g|s_{AI} \neq g]}{Pr[s_m = r|\theta = g \cap s_{AI} \neq g]Pr[\theta = g|s_{AI} \neq g] + Pr[s_m = r|\theta \neq g \cap s_{AI} \neq g]Pr[\theta \neq g|s_{AI} \neq g]} \\ =& \frac{(1-m)(1-q)(1-p)\left(\frac{n-1}{n}\right)}{qp + (1-m)(1-p)\left(\frac{n-1}{n}\right)} \in [0, \beta_b(p)].\end{aligned}$$

Note $\beta_{-g}^m(p)$ is increasing in n and $\lim_{n\to\infty} \beta_{-g}^m(p) = \frac{(1-m)(1-p)(1-q)}{(1-p)(1-m)+qp}$.

A higher number of states of the world strengthens the belief that the state of the world is good when the AI recommendation is favorable, with and without monitoring. The reason is that the observation of $s_{AI} = g$ is less likely to be due to chance; therefore, it is a more convincing signal that the state of the world is indeed good. At the same time, an unfavorable AI recommendation more weakly affects the belief that the state of the world is instead good when n is higher: if the AI system is unreliable, it is more likely to generate a signal different from $s_{AI} = g$. Despite these differences, our results would qualitatively carry over to this scenario with more states of the world.