Mitigating Generative AI Hallucinations

Alessandro De Chiara (UB), Ester Manna (UB), and Shubhranshu Singh (Johns Hopkins University)

Postal Economics Conference

April 16-17, 2026

Introduction

Promise and Peril of Generative AI

Launch of ChatGPT in 2022 stunned the world

- Generative AI generates new content, such as text, images, music, lines of software codes
- Future applications go well beyond AI chatbot assistance

1. Help in diagnosing diseases and discovering new drugs
2. Designing prototypes or even infrastructures

New regulatory challenges: Who should be liable for the harm brought about by the AI system output?

- Incorrect diagnosis, flawed design or buggy software

Introduction

AI Hallucinations

Generative AI systems may *hallucinate*

- They may provide confident responses that are in fact false

Example:

- A NY lawyer used ChatGPT in writing a legal brief
- ChatGPT invented quotes and citations from 6 cases, which were all bogus

Why? They are designed to give answers that are statistically likely

- ... answers may not be factually correct!

Introduction

Regulatory Attention

A race to update regulations

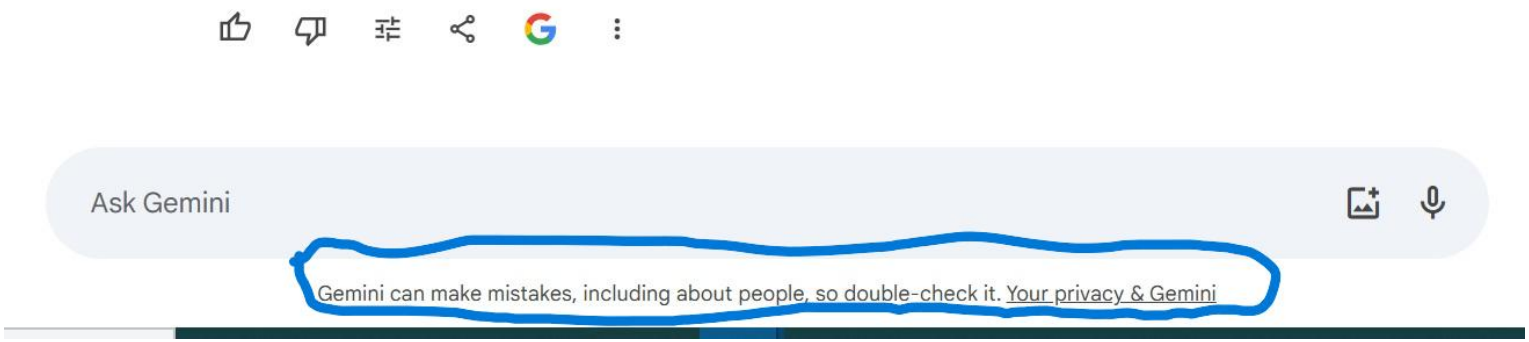
- EU at the forefront of this regulatory push
  - *AI Act* and *revised Product Liability Directive*
  - Classification of AI depending on risk level
  - Both AI developer and AI operator can be held accountable
- In the US, unclear whether *Section 230 of CDA* protects AI developers
- Critics argue that liability may *chill innovation*

Introduction

Measures Taken by Firms

In the meantime, AI developers are taking precautions

Google's Gemini:



Introduction

Copyright Issues

Related issue: copyright infringements

- Gen AI could compose texts, music, computer lines, or create images that do not sufficiently depart from copyright-protected material

Microsoft's Customer Copyright Commitment:

> ***01.05.2024 Update:*** *On November 15, 2023, Microsoft announced the expansion of the Copilot Copyright Commitment, now called the Customer Copyright Commitment, to include commercial customers using the Azure OpenAI Service. By extending the Commitment to cover the outputs from the Azure OpenAI Service, Microsoft is broadening our commitment to defend these customers and pay for any adverse judgments if they are sued for copyright infringement for the use of the Azure OpenAI Service outputs. This expansion of our copyright commitment is intended to further address customer concerns relating to potential IP infringement liability that could result from the use of the output of Microsoft's Copilots and Azure OpenAI Service. Our customers must have implemented the required guardrails and mitigations we have made available to be eligible for the benefits provided by the Customer Copyright Commitment. For our Azure OpenAI Service, we offer documentation and tooling that support the responsible use of AI and reduce risks of infringing copyrighted content.*

Introduction

Our Model

Theoretical model where:

1. An AI developer can invest to improve accuracy of AI
2. A human decision maker decides
   - 2.1 whether to use Gen AI to make a decision
   - 2.2 whether to double check the reliability of AI output before adopting it
3. Non-modeled third parties who can be harmed by the decision

Set Up

Human Decision Maker

A human decision maker (DM) chooses an activity a that can be either

- safe $a = s \Rightarrow$ no gain, no harm or
- risky $a = x \Rightarrow$ private benefits, but harmful to non-modeled third parties if state of the world $\theta = b$
  - If $a = x$ causes harm, DM suffers a non-transferable disutility $d \geq 0$
  - Think of this as a reputational loss

Before choosing a, DM can use Generative AI

Set Up

AI Developer and Generative AI

AI system analyzes existing datasets to generate a signal s_{AI} on θ

- Or, equivalently, a recommendation on which activity a to pursue

The AI system is reliable if the information used to generate the signal is exact

- The probability that the AI is reliable is $p \in [p_0, 1]$
- $p_0 \in [0, 1)$ is the publicly-known baseline accuracy
- AI developer can invest to increase accuracy

Set Up

AI Hallucinations and AI Supervision

AI can hallucinate:

- If the AI is reliable, the signal is correct ($s_{AI} = \theta$)
- If AI is unreliable, the signal is random (it can be correct by chance!)

DM can costly engage in AI supervision to double check the reliability of the AI system

- AI supervision reveals a hallucination with some positive prob.
- If DM learns that AI is unreliable, the recommendation must be discarded

Set Up

Timing of the game

Sequence of events:

1. Nature draws $\theta \in \{g, b\}$
2. AI developer invests in p and sets price T for use of the AI system
3. If DM purchases AI system, s_{AI} is generated, and DM decides whether to engage in AI supervision
4. The DM chooses $a \in \{s, x\}$ and payoffs realize

We focus on high-risk activities

Liability Regimes

We compare **liability rules** for harm caused to third parties

1. AI-Operator liability: DM is liable
2. AI-Developer liability: AI developer is liable if DM follows AI system's recommendation
3. AI-Developer Conditional liability: DM is liable unless he engages in AI supervision

We also distinguish between

(a) Ex-post efficiency: efficient selection of activity given cost of acquiring information

(b) Ex-ante efficiency: efficient provision of incentives to AI developer to invest in accuracy

Liability for the AI operator (DM)

- If $s_{AI} = b$, DM will choose the *safe* activity for any p
- If $s_{AI} = g$, DM may choose the *risky* activity
  - But first, DM must decide whether to engage in AI supervision

When does the DM engage in AI supervision after $s_{AI} = g$? If

(i) *supervision is sufficiently efficient*

(ii) and if *AI system accuracy intermediate*: $p \in [\underline{p}_h, \bar{p}_h]$

- If $p < \underline{p}_h$, DM never chooses $a = x$,
- whereas if $p > \bar{p}_h$ DM chooses $a = x$ without supervising AI

DM's incentives for AI supervision and activity selection are socially efficient

Liability for the AI developer

Akin to AI Operator liability, $s_{AI} = g$ improves the DM's belief that the state is good

- On top of this, AI allows the DM to **shift liability**

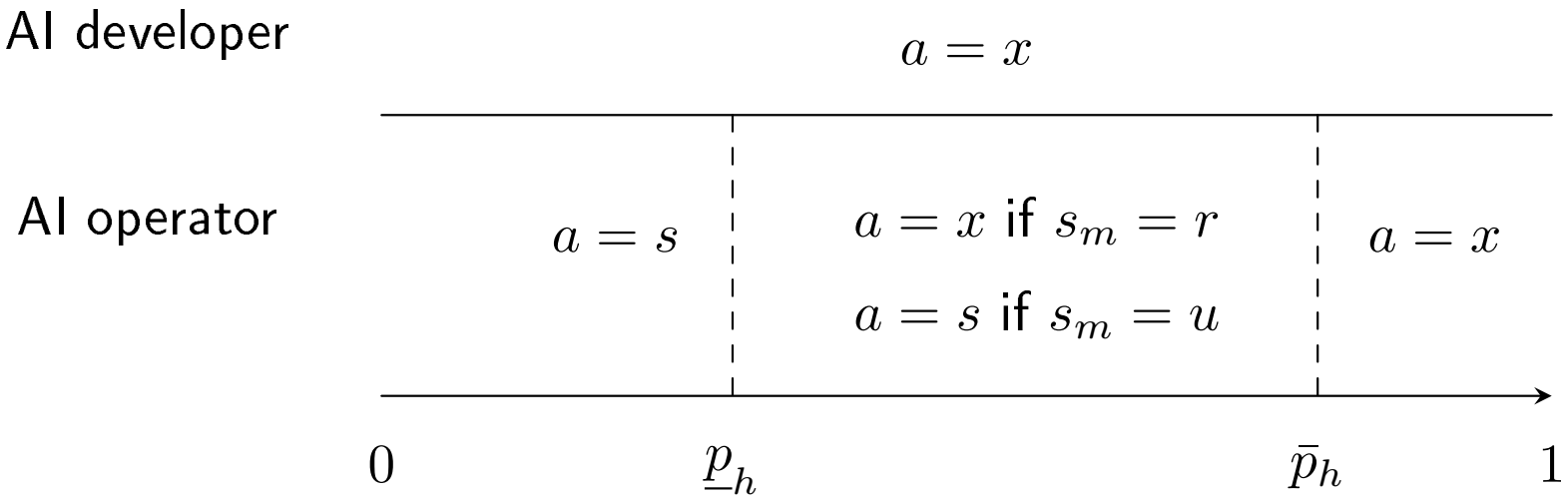
$\Rightarrow$ Social and private (DM's) incentive for AI supervision and activity selection are not aligned

- Following $s_{AI} = g$, DM only considers d when deciding whether to supervise and choose a risky activity

Yet, the *direction of the distortion* may be counterintuitive

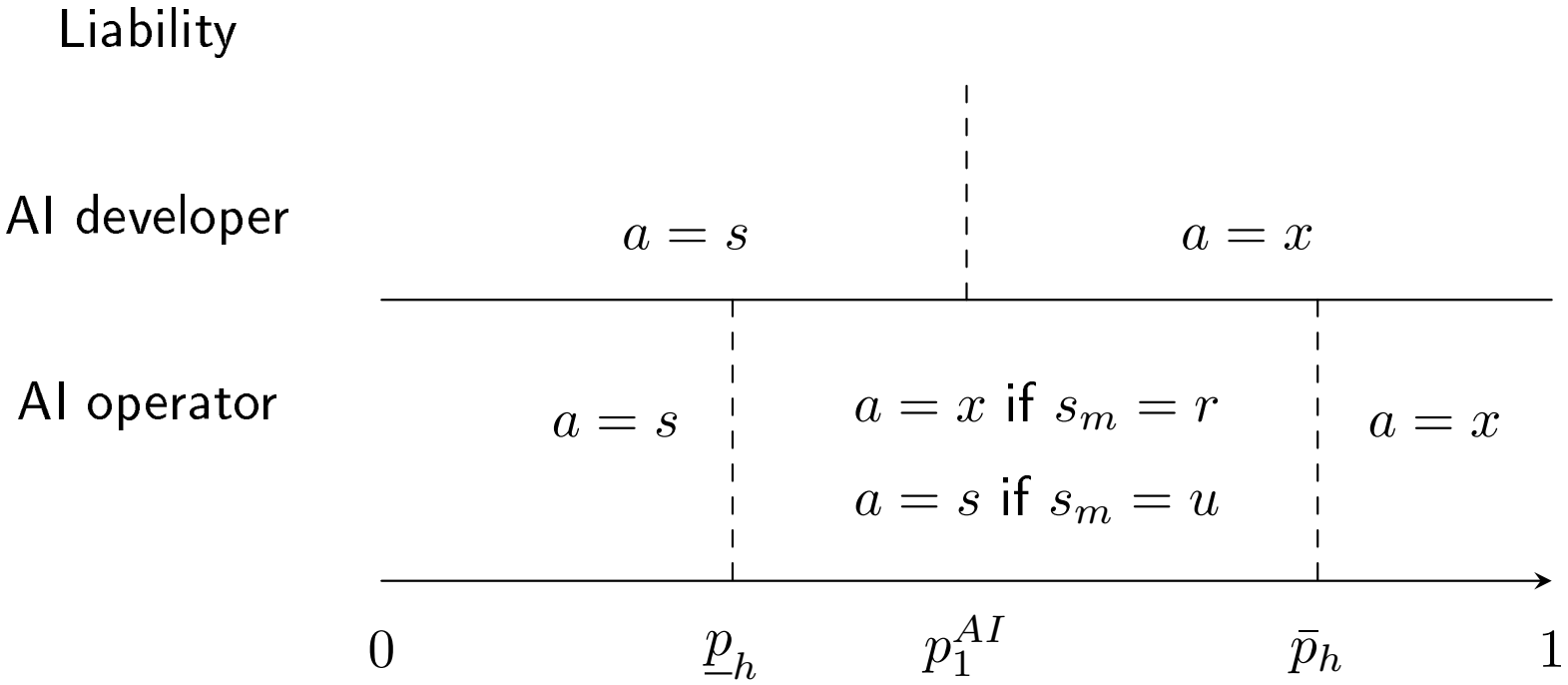
Comparison of Liability Regimes

There is no supervision and excessive risk-taking under AI Developer liability when reputational concerns are minor



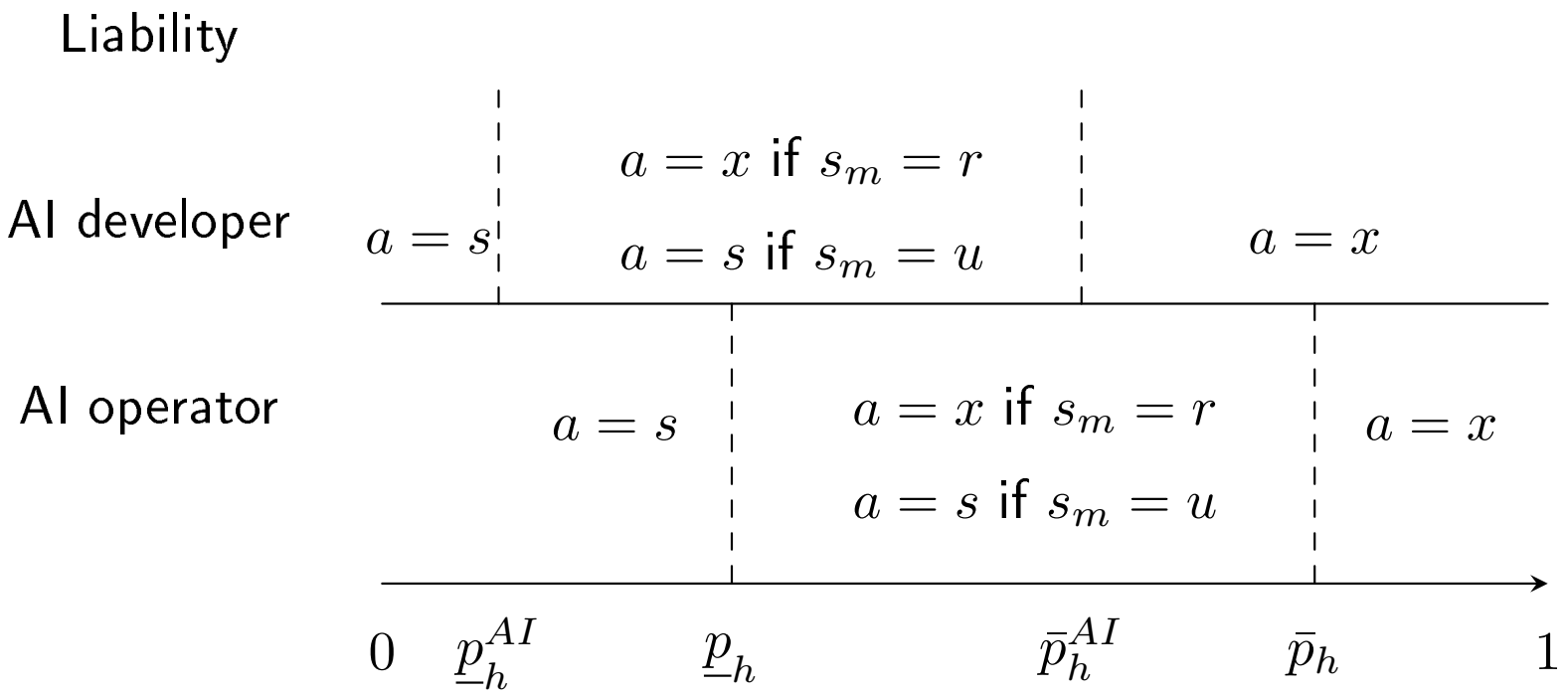
Comparison of Liability Regimes

There might be too little risk-taking under AI Developer liability when supervision is not efficient and there are significant reputational concerns



Comparison of Liability Regimes

There might be excessive AI supervision under AI Developer liability when supervision is efficient and there are significant reputational concerns



Ex-ante Efficiency

Consider AI developer's expected profit at stage 1:

$$\Pi = T - c(\Delta p, p_0) - HPr[H]\mathbb{1},$$

where

- $Pr[H]$ is the probability that harm occurs
- $\mathbb{1}$ is an indicator function
  - $\mathbb{1}$ takes value 1 if the AI developer is liable for harm to third parties
  - $\mathbb{1}$ takes value 0 otherwise

AI developer chooses $T \geq 0$ and $p \in [p_0, 1]$ to maximize Π subject to DM's acceptance of the contract

Ex-ante Efficiency

Public and Private Investment

Remark: if p is publicly observable, AI Operator liability is ex-ante and ex-post efficient

- Intuition: private and social incentives are ex-post aligned and AI developer will choose p to maximize $T(p) - c(p, p_0)$
- Result reminiscent of Hay and Spier (2006, AER)

Remark: if p is privately observable, AI Operator liability leads to no investment: $p^{DM} = p_0$

- Intuition: investment cannot directly influence T

$\Rightarrow$ If investment is private, AI Developer liability may be desirable

- AI developer invests to reduce expected liability cost

Ex-ante Efficiency

Private Investment

Proposition: AI Operator liability is desirable only

(i) In the parameter region $[\underline{p}_h, \bar{p}_h]$

(ii) If it is the only regime that encourages AI supervision

Intuition:

- AI developer can refuse to sell AI system to DM $\Rightarrow$ This attenuates ex-post distortions
- AI developer has incentive to invest if she anticipates liability
- Supervising AI system may be desirable

Ex-ante Efficiency

Private Investment

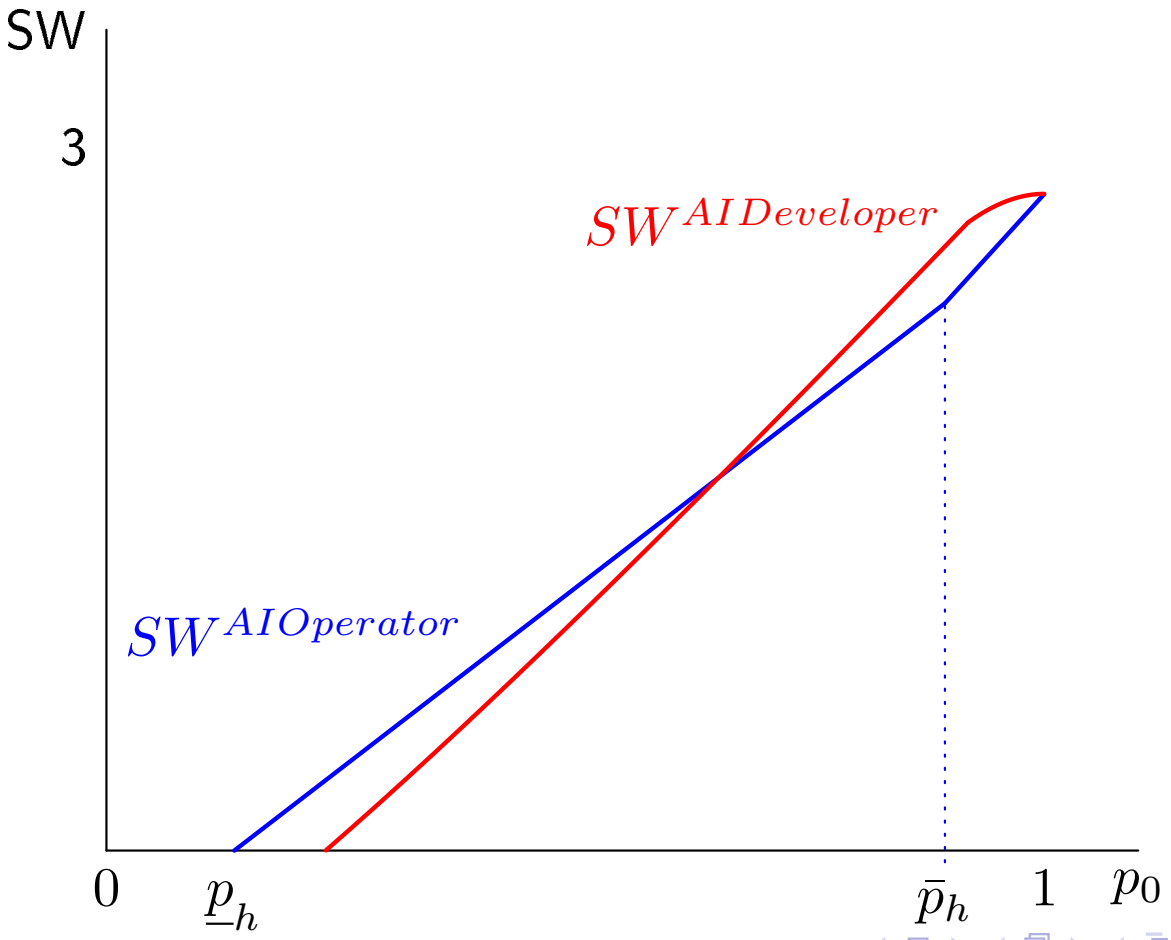
Proposition Cont'd: There exists $\tilde{p}_0 < 1$ such that for any $p_0 > \tilde{p}_0$ AI Developer liability results in higher welfare

- For p_0 low, AI system is not adopted
- For p_0 intermediate, AI can be adopted and AI Operator liability may dominate
- For p_0 sufficiently high, AI Developer liability always better
  - No AI supervision under either regime
  - AI Dev liability leads to more accuracy

If p_0 captures the quality of the baseline tech, optimal liability regime may change over time

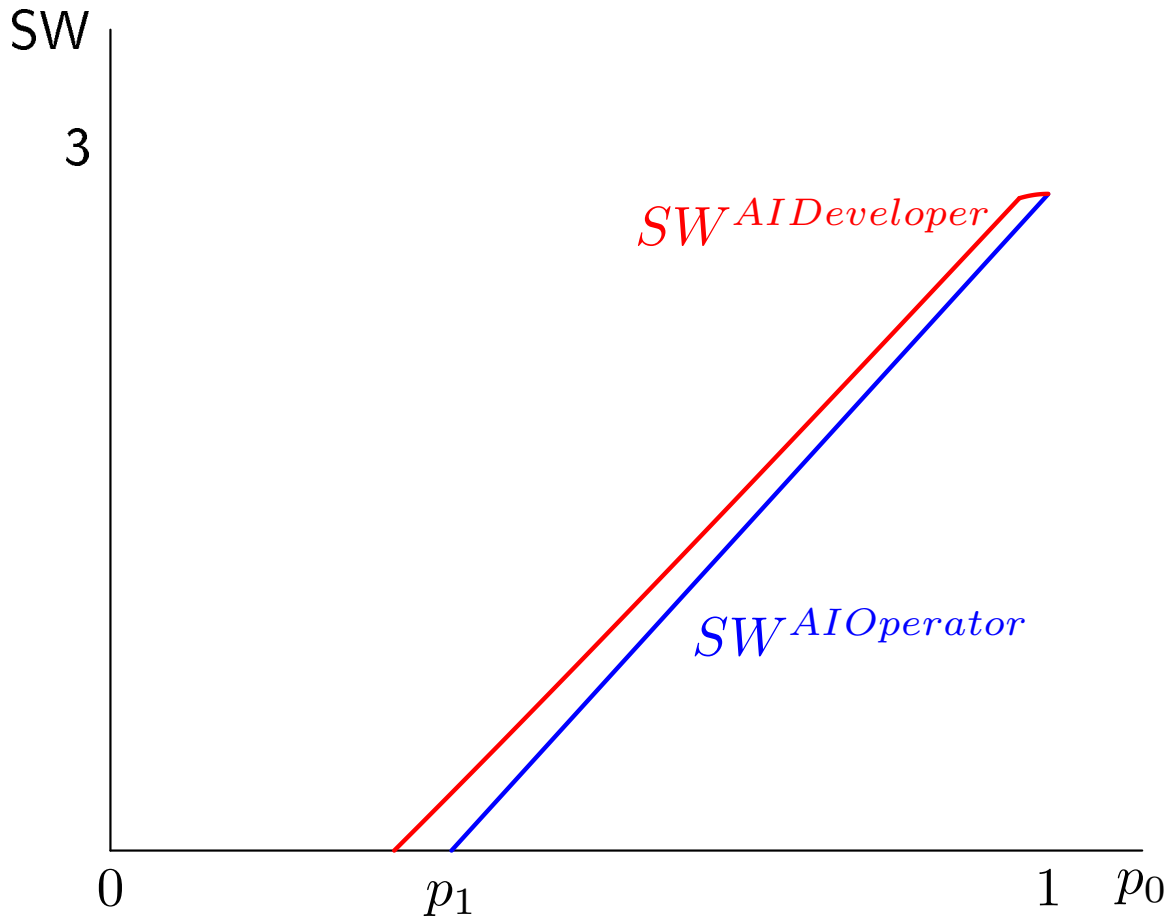
Ex-ante Efficiency - Graphical Illustration

AI Operator liability may dominate for intermediate p_0 when it is the only regime that induces AI supervision:



Ex-ante Efficiency - Graphical Illustration

AI Operator liability never dominates when it is not the only regime that induces AI supervision:



AI-Developer Conditional Liability

Alternative liability regime:

- DM liable if he did not supervise AI
- AI developer liable if DM supervised AI

Pros:

- It strengthens incentives to supervise compared to AI Dev liability
- It fosters incentives to invest compared to AI Oper liability

Cons:

- AI supervision can be excessive (*liability shifting*)
- For p_0 high enough, AI Dev liability always dominates
- Can we *verify* that AI supervision took place? Or its result?

Low-risk activities

Low-risk activities

- When $s_{AI} = g$, the DM always selects the risky activity irrespective of p and of the allocation of liability
- When $s_{AI} = b$, the DM may decide to choose the risky activity, but cannot shift liability if things go awry
- There is no distortion in the choice activity under AI developer liability

AI developer liability dominates

Conclusion

Framework to study how liability affects interaction between Gen AI and human decision makers

- Gen AI critical for DM's actions but may be unreliable
- Liability affects how it is used by the DM and AI developer incentives
- Riskiness of decisions and whether accuracy of technology is observed matter
- and so does the DM's reputation
- Efficient liability regime may change as technology evolves

Paper may help rationalize some business strategies

Related Literature

Liability & AI Innovation: We join the discussion among legal scholars and economists (Galasso and Luo, 2017, 2022, Buiten, 2024, Chen and Hua, 2026)

- focusing on how liability allocation between AI operators and developers affects AI adoption and AI precision.

Principal-Agent Models & AI: We relate to models exploring delegation to AI or human agents and AI's impact on strategic communication (Athey et al., 2020, Gans, 2024, Llanes and Madio, 2024)

- We consider an agent's decision to use and follow AI for decisions that may cause harm where users do not observe AI accuracy

Liability & Product Safety: We adapt traditional product liability literature to AI, considering its specificities and interplay with human decisions