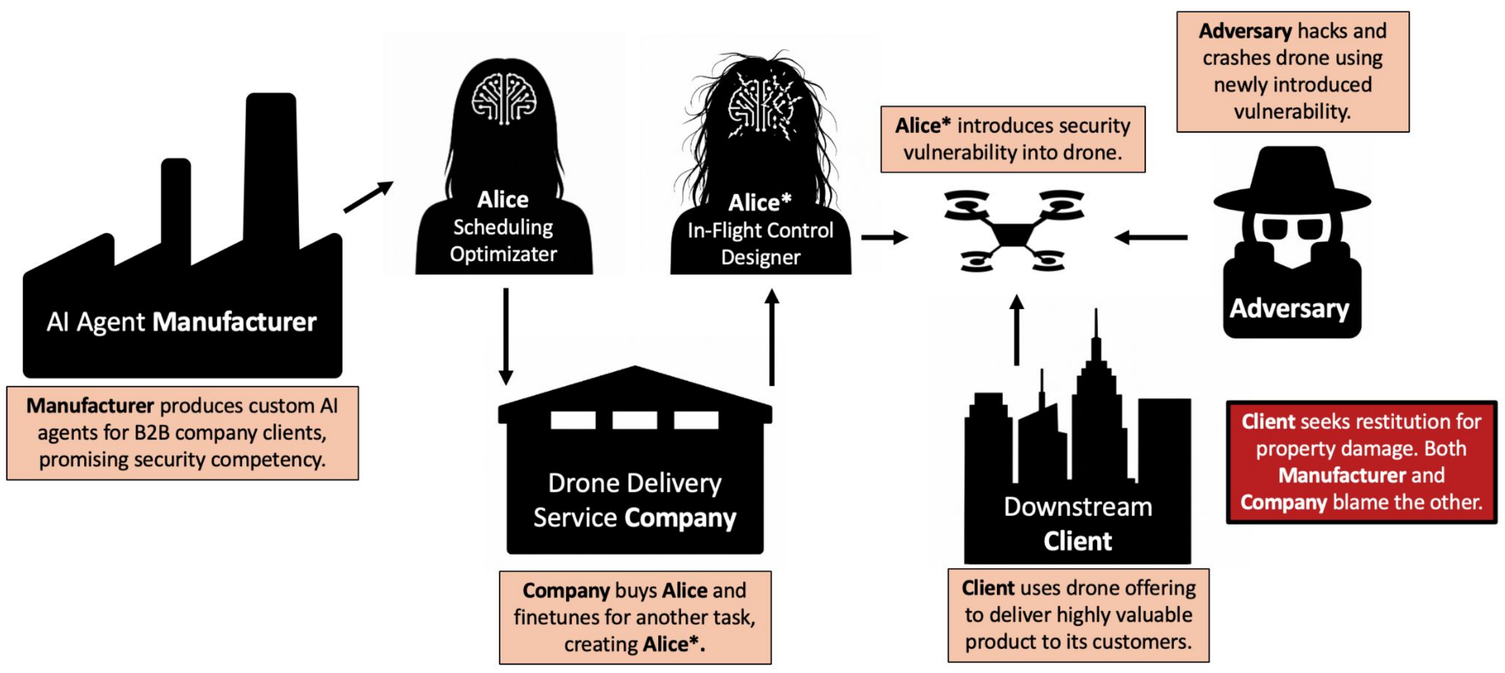
# Bridging the Gap Between Tort Law and Unforeseeable AI Errors

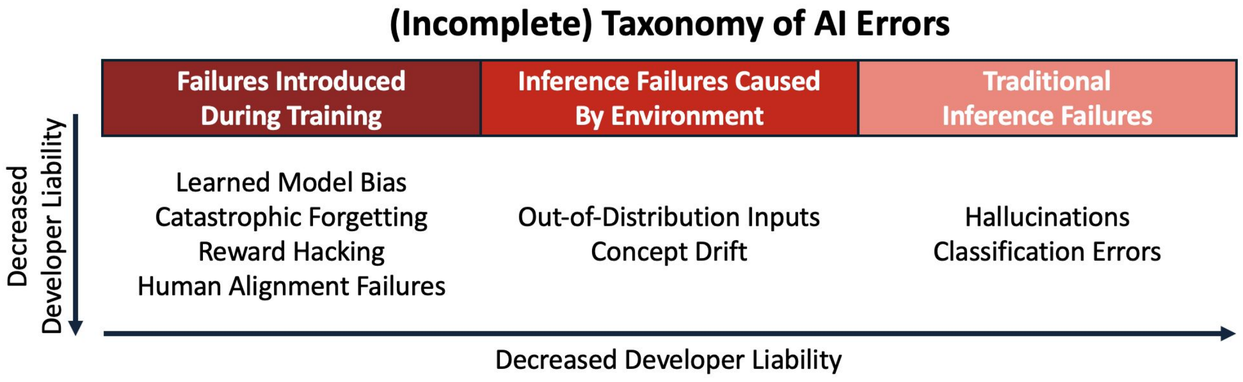
Mitchell Chervu Johnston
Boston College Law School

Brian Matejek
SRI International



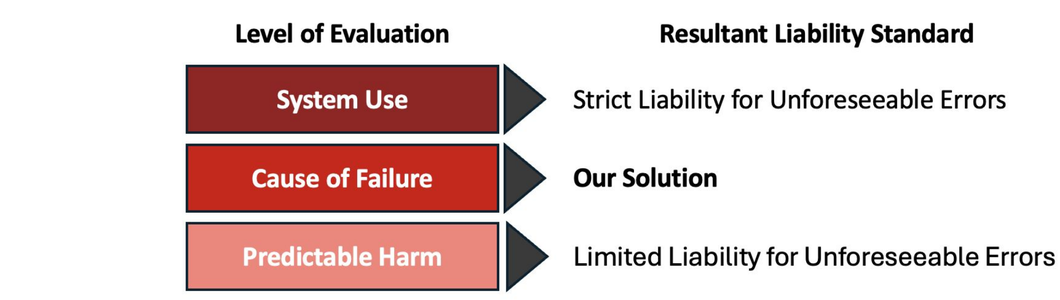

- Tort law has many general categories of liability that can be applied to novel technologies like AI.
- Liability can be strict (defendant compensates plaintiff independent of fault) or fault-based (plaintiff must show that defendant failed to meet the legal standard of appropriate conduct).
- We believe true strict liability is unattractive both for fairness reasons and due to its ambiguous deterrence effects.
- Attempts to incorporate unpredictable errors into traditional doctrines results in *de facto* strict liability as developers and users of AI have no ability to prevent or control these errors.
- We propose that liability for errors should be based on the *ex post knowability* of unusual errors as a substitute for foreseeability.
- This solution balances the goal of fairness and positive incentives for developers with the goal of compensating injured parties.
- Our ideas can be applied beyond US law if lawmakers wish to adopt a flexible standard of liability based on the principles we articulate to complement regulation.

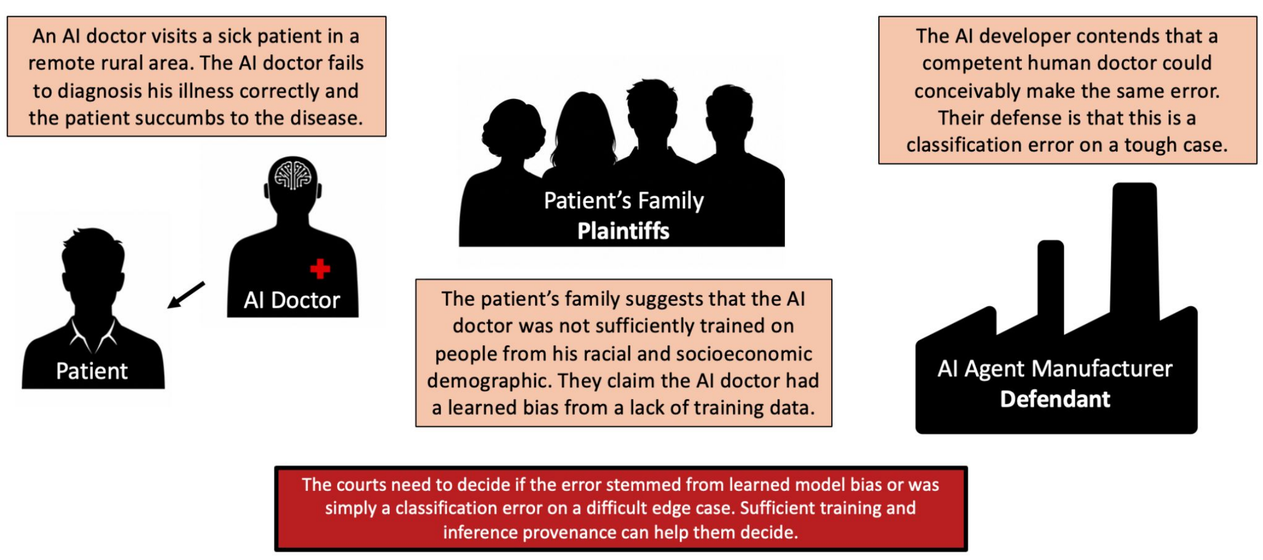
## (Incomplete) Taxonomy of AI Errors




- Liability is not governed fully by the type of error and category. The environment surrounding the error can significantly influence liability.
- Consider a "Level 2" semi-autonomous vehicle, such as a Tesla, that requires full human-supervision over the vehicle when in "self-driving" mode.
- Imagine three unique scenarios with similar errors: the car bypasses a stopped school bus, ignoring the flashing stop sign, and hits a disembarking student; a woman turns on self-driving mode in white-out winter conditions and crashes into a parked car; and an influencer turns on self-driving mode and moves to the backseat before the car drifts into oncoming traffic.
- AI developer liability would decrease in each scenario as the user abuses the capabilities of the system, assuming sufficient notice of the limitations of self-driving mode and the necessity of a driver-in-the-loop.

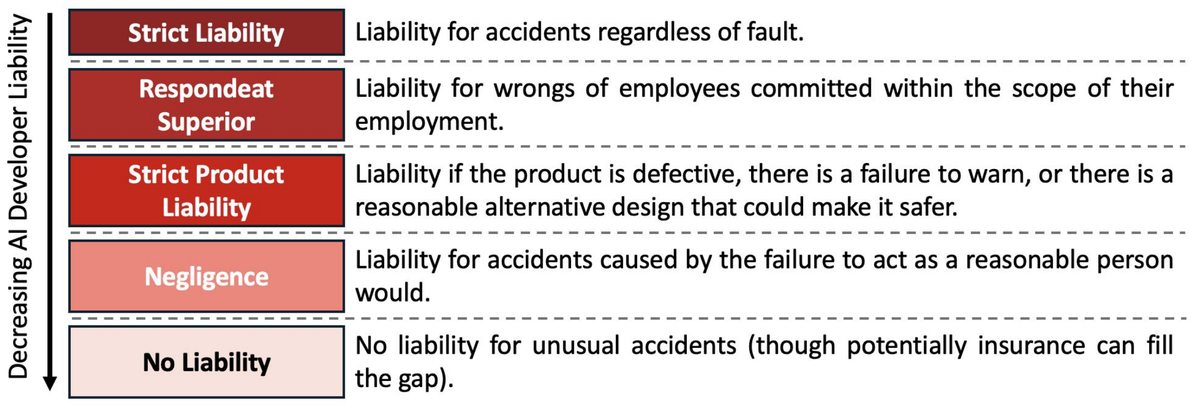
## Potential Levels of Generality for Liability




- Tort law provides a remedy for individuals harmed (including accidentally) by the actions of others.
- Tort law (we argue) relies on social understandings of how harms occur to guide behavior and attribute fault.
- AI systems can err in *unpredictable* and *unusual* ways that are not socially understood.
- AI developers need to produce AI systems that enable the courts to fairly adjudicate the causes of the failures.
- We envision *provenance* measures to allow courts to reconstruct the root causes of errors.
- *Data provenance* could help identify deficiencies, such as learned model bias, during training.
- *Inference provenance* could help identify the cause of an error, for example, degraded decision making in an evolving world (e.g., concept drift).

## Traditional Tort Law Categories

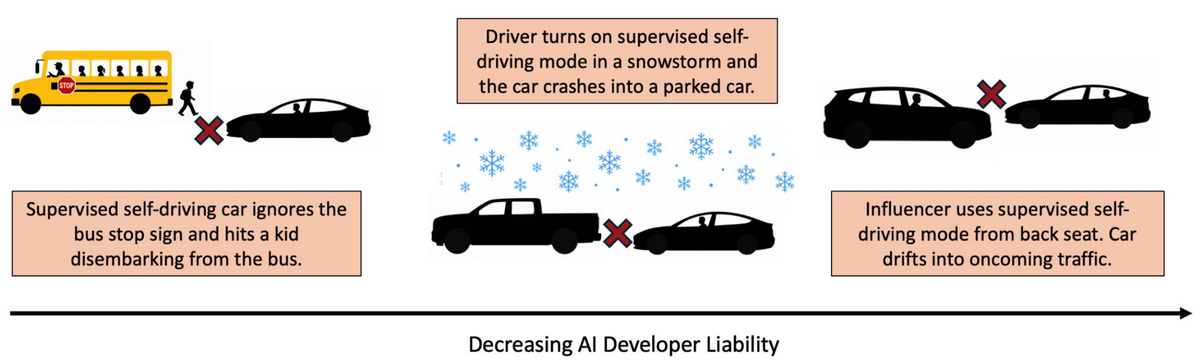



## Proposed Liability Framework

| | No Notice | Sufficient Notice |
|---|---|---|
| Known Failure Cause | **Potential Developer Responsibility** | **Potential Mixed Responsibility** |
| Unknown Failure Cause | **No Liability** | **User Responsibility** |

- There are a large number of known ways that AI systems fail.
- These errors can arise through failures during training (e.g., learned model bias or reward hacking) or lack of safeguards during inference (e.g., out-of-distribution errors and hallucinations).
- The potential mixed responsibility depends heavily on the *type of error* of the AI system.
- Training errors typically shifts liability towards the AI developer with less liability during inference. For example, a standard classification error, particularly one that a reasonable human might make, could have limited to no liability implications for the developer.

## AI Failures and Environmental Context




## Governing Principles

- **Principle 1:** Errors that cannot be traced to a known cause of AI failure should not be the basis for liability. Alternatively stated, traceability is a necessary, but not sufficient condition for liability.
- **Principle 2:** For errors that are traceable, researchers and courts will need to develop appropriate standards for how developers should attempt to prevent harmful errors. And, in some cases, meeting those professional standards should constitute a defense to liability. The relevant professional standard may also vary depending on whether the error is emergent or the result of deliberate design choices.
- **Principle 3:** Where an AI system is deployed by an external party in a way that harms a third party, liability rules should take into account which actor was in the best position to know about the potential for error and to prevent the harm from occurring. In particular, in the case of specialized actors (e.g., doctors) using AI systems, sufficient warnings about the potential for and nature of errors should typically suffice to shift responsibility to the specialized actor.
- **Principle 4:** AI developers need to estable training and inference provenance where possible into their systems so that the courts can adequately determine the cause of an AI system failure. Plaintiffs and defendants will often offer competing causes of the failure that could shift liability from one party to the other.