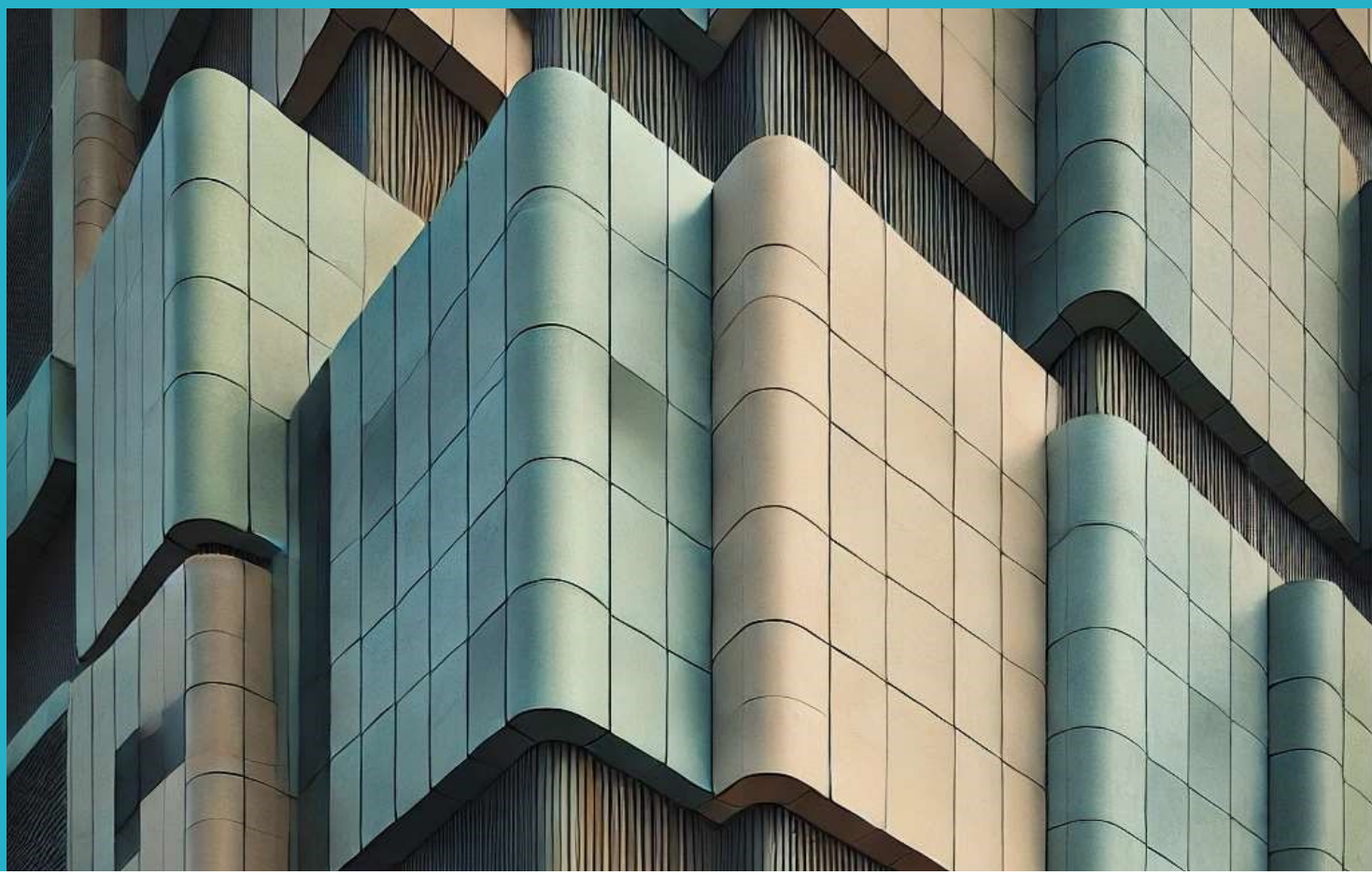
WHITEPAPER
November 2024

INSURING EMERGING RISKS FROM AI

Authors: Gabriel Weil, Matteo Pistillo, Suzanne Van Arsdale, Junichi Ikegami, Kensuke Onuma, Megumi Okawa and Michael A. Osborne



In partnership with

INSURING EMERGING RISKS FROM AI

Gabriel Weil[1,2], Matteo Pistillo[2], Suzanne Van Arsdale[2], Junichi Ikegami[3], Kensuke Onuma[3], Megumi Okawa[4] and Michael A. Osborne[5]

14 November 2024
Correspondence to Michael Osborne, mosb@robots.ox.ac.uk

[1] Touro University Jacob D. Fuchsberg Law Center
[2] Institute for Law & AI, law-ai.org
[3] Aioi R&D Lab Ltd.
[4] Aioi Nissay Dowa Insurance Co. Ltd.
[5] Oxford Martin AI Governance Initiative, University of Oxford

Contents

Abstract

This report examines the implications of recent progress in artificial intelligence (AI) for liability regimes and insurance markets within the United States. We argue that the insurance industry faces both a potential decline in traditional markets like auto insurance and emerging growth opportunities in AI agent and cybersecurity coverage. The report advocates for targeted reforms in liability laws, proposing a nuanced approach that may ease regulations for demonstrably-safer technologies, such as future autonomous vehicles, whilst strengthening oversight for AI agents and cyber risks. Key recommendations include implementing strict liability regimes for a subset of AI harms, mandating insurance coverage for certain AI applications, and expanding punitive damages to address catastrophic, uninsurable risks. These proposed changes would significantly impact the insurance sector, necessitating the development of new actuarial methodologies to quantify complex AI-related risks and to potentially underwrite a broader range of liabilities. We conclude that the insurance industry has a pivotal role to play in managing AI-related risks, fostering responsible innovation, and ensuring that the benefits of AI are broadly shared across society.

Introduction

The past decade has seen rapid advances in AI technology that are only now beginning to filter out into the broader economy. AI systems have learned how to predict protein structures, compose original poetry and essays, beat the world's best players of Go and Starcraft, generate pictures, music, and video from text prompts, hold conversations, and score impressively on a range of standardized tests. While progress has been slower in embodied systems, autonomous taxis have been deployed in several major cities. Future AI systems may hold the potential to supercharge economic growth and innovation, help cure major diseases, and transform the ways that most people live and work. With broad deployment, autonomous vehicles (AVs) could greatly increase mobility (especially for the elderly, children, and disabled), and enable better use of urban land. But AI-powered automation also poses substantial risks of harm to users and third parties alike.

Insurance can help ensure that AI-related harms are mitigated, and that AI's risks and benefits are fairly shared. Insurance offers several tools for achieving these goals. First, insurance can share risk, narrowing the gap between the winners and losers from AI's broad adoption. Second, insurance can promote AI's benefits, via permitting firms that are too small to self-insure to take on uncertain AI investments. Insurance can also induce more responsible usage of the technology. For instance, auto insurance can lead people to drive more safely—auto insurers commonly offer both advice to improve driving behaviour and premium discounts to incentivize safe driving. When it comes to insurance for AI, by offering risk-based premiums tied to safety standards and certification, insurance can create financial incentives for developers to prioritize responsible AI development. Furthermore, mandatory AI insurance for developers or deployers could ensure a minimum level of financial protection for potential victims of AI-related incidents, similar to the function of mandatory auto insurance for drivers.

In many industries, demand for insurance is driven by the structure of legal liability rules, which often make one party responsible for paying for injuries suffered by another. Liability, like insurance itself, is an important tool that our society has developed for managing the risks citizens expose to one another, even when engaged in broadly-beneficial activities. Four forms of liability are most relevant to the insurance markets analyzed in this report:

Negligence law imposes a general duty to exercise reasonable care to prevent foreseeable physical injuries.

Products liability holds commercial sellers of products strictly liable for injuries resulting from the foreseeable uses of products that are defective in their design, manufacturing, or accompanying information.

The *abnormally dangerous activities doctrine* holds individuals and corporations strictly liable for any foreseeable harms caused when they engage in unusual activities that generate a high risk of harm even when reasonable care is exercised.

Finally, the doctrine of *respondeat superior* holds employers vicariously liable for the torts committed by their employees, within the scope of their employment. This doctrine could plausibly be extended to AI agents that act on behalf of designers or users.

Deployments of AI systems are likely to trigger, under different circumstances, each of these four sources of liability. The rollout of AI systems may also shift which of these regimes apply in specific domains. For instance, the law that currently applies to most auto collisions is negligence, applied to the driver's conduct. Sometimes products liability is triggered by evidence that a vehicle involved was defective in its design, manufacture, or warnings, but the analysis typically begins with the care exercised by the driver. As more driving functions are automated, the locus of analysis will likely shift away from the care exercised by the driver to the reasonableness of the design of the vehicle, including the algorithms that enable its autonomous capabilities.

These changes will likely also transform insurance markets. As driver negligence recedes as a source of liability, the case for requiring individual road users to carry liability insurance will weaken. Indeed, in AVs that are fully responsible for all driving tasks, like Waymo taxis, it is already the case that no licensed and insured driver need be present in the vehicle when it is operating. Instead the liability burden, and possibly mandates or economic incentives spurring demand for insurance, will shift to the vehicle manufacturers and the software companies designing the AV algorithms.

While the benefits and risks of AI are likely to pervade broad swathes of the economy, this report focuses on three specific domains: AVs, AI agents, and cybersecurity. AVs are classified into six levels of automation, but generally involve delegating some or all driving tasks to an automated system. AI agents are an emerging form of AI that acts autonomously to accomplish goals provided by the system user. The scope of the user-specified goals can vary widely; this category includes systems that range from chatbots, to copilots, to advanced AI assistants in the form of digital or robotic systems that can execute complex workflows autonomously. Finally, AI is likely to heighten both the importance of cybersecurity and the risks of cyberattacks, including risks of financial loss, disruption or reputational damage.

This report analyzes the potential role of insurance in managing the emerging risk associated with progress within AI. While insurance will likely have a substantial role to play in many countries, each with different legal systems, in this report, we focus on U.S. liability law, the US being both the home of the largest AI firms and the largest insurance market in the world. We provide an analysis of how this body of law may treat future AI risks, and discuss what the potential consequences may be for insurance markets.

We argue that AI-driven automation may lead to declining demand in sectors like auto insurance, but will offer growth opportunities in AI agent and cybersecurity coverage. This shift suggests a strong case for reforms in liability laws, potentially easing regulations for safer technologies like autonomous vehicles, whilst strengthening regulations for AI agents and cyber risks. We advocate for strict liability for certain AI harms, insurance mandates, and expanded punitive damages to address uninsurable catastrophic risks. These changes would significantly impact the insurance industry, requiring insurers to adapt by quantifying complex AI-related risks and potentially underwriting a broader range of liabilities, including those stemming from "near miss" scenarios. AI itself may provide new affordances for insurers in modelling the risks of mission-critical AI system. In short, insurance has an essential but demanding role to play in the future of AI.

Overview of AI Risks and Affected Entities

At the highest level of generality, there are three categories of failure that can lead to AI harms.

First, the system may cause harm because of a *capabilities failure*: its capabilities fall short of those demanded by the deployment context. Capabilities failure is the most likely cause of harms from AVs, and may also be a substantial source of harms caused by AI agents. With the long-standing trend of progress in capabilities, capabilities failures are likely to decrease in frequency and severity over time.[6] However, if the number of frontier AI developers remains limited, the risks of capabilities failures may be correlated—many applications may suffer from the failure of a single model.

Second, AI systems may cause harm because they are *misaligned*. That is, an AI-powered system may be capable of completing the tasks assigned by the user, but may have conflicting goals from the user's. This sort of alignment failure is most likely to arise in the context of AI agents, but could also arise for AVs. In contrast to capabilities failure, the risks associated with alignment failures may actually rise over time, as advances in AI capabilities, which have historically outpaced advances in AI alignment, increase the scale and severity of the harms that misaligned systems cause.

Finally, an AI system might be *misused* by a user who instructs the systems to cause harm.[7] This sort of AI misuse might happen with any AI-powered system, though it seems unlikely for AVs, given their limited user input channels, at least in prevailing designs. Cybersecurity breaches are a particularly prominent vector for AI misuse. This includes scenarios in which AI systems are used to identify and exploit cybersecurity vulnerabilities as well as cybersecurity breaches that give malicious users access to powerful AI systems that they then use to do harm.

For capabilities failures and alignment failures, any liability will generally tend to fall on the developers and providers of AI systems. For example, manufacturers and sellers of AVs, and possibly also the developers of the AV algorithms, are likely to have liability exposure when AVs are involved in collisions. The liability of the algorithm developers is likely to depend on the contractual relationship between automakers and their software providers as well as the classification of AV algorithms as a product or a service. Users may also be liable for capabilities failures if they know or should know the

6 Capabilities failures could also increase if the scale of automation and complexity of the tasks demanded of AI systems outpaces the advances in capabilities.

7 This instruction might be given with the intent to cause harm, or might merely instruct the system that take actions that generate unreasonable risk of harm. The former would be an intentional tort on the part of the user, while the latter fall under negligence law.

limits of a system's capabilities and deploy the system in a manner that is unreasonably dangerous in light of those limits.

For harms associated with misuse of AI systems, the users will be liable. But users of AI systems may be effectively judgment-proof. For example, there is concern that advances in AI technology will make it easier for terrorist groups to construct chemical, biological, radiological, or nuclear (CBRN) weapons. While any terrorist group that launched a CBRN attack would be subject to civil (not to mention criminal) liability, terrorists are unlikely to be deterred by this prospect, as they are unlikely to have sufficient assets subjects to the jurisdiction of the relevant courts to pay out a substantial portion of any damages award. This raises the question of whether the developers or providers of AI systems that are susceptible to such misuse may also be held liable. The legal regimes under which developers, providers, and users of AI systems might be held liable will be discussed in detail below.

The shifts in liability induced by AI are likely to have significant implications for insurance markets. For instance, as AV deployment scales up, liability for auto collisions is likely to shift from driver negligence to manufacturer products liability. This is likely to decrease demand for auto insurance, since manufacturers will be better positioned to self-insure against liability risk than individual drivers.

By contrast, AI agents and cybersecurity concerns present larger risks that may generate new demand from AI developers for liability insurance. These larger-scale risks might even push policymakers to require AI developers and providers to take out liability insurance or otherwise demonstrate the ability to pay out damages awards commensurate with the potential harms of their systems. Insuring against large scale risks could be a major growth market in the coming years. In fact, the catastrophic risks about which many prominent experts warn may push the limits of insurability, threatening damage measures in the hundreds of billions of dollars (e.g. the failure of an AI system controlling critical infrastructure). The maximum size of insurable risks may emerge as a central question of insurers and policymakers alike in the coming years.

Towards Full Automation: Societal Benefits and Implications for Liability and Insurance

AI is enabling new forms of automation that may bring significant societal benefits. Consider, for instance, three applications of advanced AI: autonomous vehicles, AI agents, and cyberattacks.

Autonomous Vehicles

When applied to vehicles, advanced AI is enabling the evolution of driving from Levels 2 and 3 (semi-autonomous driving) to Levels 4 and 5 (fully-autonomous driving). The levels of driving automation refer to the Society of Automotive Engineers (SAE) classification of driving automation in five different levels, which have been adopted by the Department of Transportation, National Highway Traffic Safety Administration (NHTSA). Levels span from Level 0—which describes only a vehicle's "momentary assistance" to the driver—to Level 5—which describes a condition of "full automation," i.e., when the car is "fully responsible for all driving tasks while any occupants act as passengers and do not need to be engaged." Intermediate Levels 2, 3, and 4 correspond, respectively, to: (2) "additional assistance"—when the vehicle "provides continuous assistance with both acceleration/braking and steering, while driver remains fully engaged and attentive," such as in the case of Tesla's Autopilot; (3) "conditional automation"—"performs all driving tasks while the driver remains available to take over any or all tasks if prompted;" (4) "high automation"—the vehicle is "fully responsible for all driving tasks . . . but can only operate within limited service areas." It is estimated that Level 5 will be reached by 2035, though past predictions regarding advances in AV technology (including by significant players like Nissan and Toyota) have been over-optimistic, and uncertainty remains high, particularly around adoption timelines. In the meantime, autonomous vehicles are being adopted in more contexts and at larger scales. Companies including Waymo, Cruise, Zoox, and Monet Technologies are expanding the areas of coverage for their robo-taxi fleets.

Autonomous vehicles promise to substantially reduce the number of car accidents and associated fatalities, injuries and economic costs—especially as we transition to autonomous vehicles forming the majority on our streets. In the US, vehicle collisions cause roughly 5.2 million injuries and 40,000 deaths each year. Global incident rates are far higher, as road traffic crashes kill 1.19 million people each year, and injure 20 to 50 million more.

Most incidents are caused by driver error, which includes inattention, decision errors such as driving too fast for the conditions, and performance errors, among others. Besides safety, autonomous vehicles can also improve efficiency, and bolster access to transportation, by providing mobility options to people who are unable to drive, including aging populations and disabled people.

AI Agents

An AI agent is an AI system that can autonomously plan and take actions to achieve user-specified goals. The technology is nascent and currently has very limited commercial impact. While it is difficult to arrive at a fully satisfactory definition of AI agency, AI agents are distinguished from AI tools (like existing unassisted large language models) by the degree to which they act directly in the world to achieve long-horizon goals, with little human intervention or specification of how to do so.

AI agents hold the potential to automate a broad range of routine commercial tasks, including generating ideas and content; conducting market research and managing entire sales pipelines, including negotiating and automating purchases; analyzing data and writing code; to serving as a tutor, friend, confidant, coach or personal assistant.

AI agents also hold transformative potential for advancing human flourishing. Automating large chunks of the scientific discovery process could yield rapid advances in science and technology, including new treatments and cures for a wide range of diseases. Automating many job functions could free up human time and attention to focus on a narrower range of tasks on which humans still outperform AI agents or where human involvement is highly valued for other reasons.

But these potential benefits of AI agents also come with attendant risks. Since AI agents directly interact with the world, there is a greater risk of any capabilities failure, alignment failure, or misuse producing substantial harm. For instance, AI agents could plan and autonomously execute offensive cyber operations, through the identification of system vulnerabilities and malicious code generation. Misaligned AI agents could manipulate, deceive, coerce, or exploit their users because of their integration in multiple aspects of a person's life.

The initial commercial model for AI agent deployment will likely involve AI agents built on top of a small number of general-purpose AI systems (such as large language models) developed by a small number of model developers. Many intermediary firms will then procure their AI agent labor from one of these handful of model developers. This creates a number of structural risks. For instance, agents acting on behalf of competing vendors

could collude to harm consumers. The reliance of many intermediary firms on a small number of upstream model developers could also create correlated risks of failure, similar to the dynamics observed in the CrowdStrike outage earlier this year.

As AI agents improve, it may also be more competitive to delegate tasks currently done by humans to AI agents, with a subsequent risk of overreliance. That is, as the risk from capabilities failures declines, there will be competitive pressures to rely more on AI agents and less on genuine human discretion. This may leave us exposed to large tail risks from both misaligned AI agents and misuse of powerful agentic systems.

Cyberattacks

A cyberattack is a malicious and deliberate attempt to breach the information systems of another in order to "change, destroy, or steal data, as well as exploit or harm a network." In the coming years, AI systems are likely to be centrally involved in cybersecurity, both as the target of cyberattacks and as the instrumentality of such attacks. AI tools are also likely to be useful in bolstering cyberdefense.

In particular, AI can increase the accessibility, frequency, and destructiveness of cyberattacks, by lowering the barrier to entry and by increasing the "success rate, scale, speed, stealth, and potency" of such attacks. AI tools can facilitate both the identification and exploitation of systems vulnerabilities. The automation of cyberattacks via AI systems can also allow them to run in parallel and at greatly reduced cost.

Similarly, AI systems are likely to be attractive targets for cyberattacks. Current frontier AI systems cost tens of millions of dollars to train. As the capabilities of AI systems increase and AI is entrusted with the automation of important social, economic, and governmental functions, the incentives to obtain such systems will rise accordingly, and exfiltrating such systems may be easier than retraining them from scratch. In some cases, cyberattackers may merely wish to gain access to powerful systems, which they might subsequently misuse. In others, the goal of a cyberattack may be to imperil or distort the functioning of the larger systems in which AI algorithms take on a central role. It is also conceivable that attackers will extort model developers by threatening to openly publish the model weights underlying such systems, which would erode the developer's competitive edge.

Advances in AI capabilities will also provide useful tools for defending against cyberattacks. For example, defenders can also use vulnerability discovery tools to identify where patches are needed. Accordingly, the net effect of AI on the offense–defense balance is indeterminate. In any case,

liability and insurance will have important roles to play in managing the transformed risk landscape and providing incentives to secure AI systems themselves and other critical infrastructure against AI-enabled cyberattacks.

Shifting Liability Regimes and Insurance Coverage

As automation enabled by AI becomes widespread, humans will become less central to decision-making and execution. Decision-making becoming increasingly algorithmic may have important consequences for both liability regimes and insurance markets.

Under current law, the two forms of liability most relevant to AI-driven automation are negligence and products liability. For both, the plaintiff must prove that the harm suffered was a foreseeable consequence of the defendants allegedly tortious conduct. In products liability, misuse is actually a term of art that means unforeseeable use, so commercial sellers are never liable in misuse cases. But the term misuse is deployed much more broadly across analyses of AI risks, and therein does include many cases of malicious use that are indeed (if only at a sufficiently high level of generality) foreseeable.

In negligence cases, the plaintiff must also show that the defendant failed to adopt some precautionary measures that a reasonable person would have adopted, and that would have prevented the plaintiff's injury. For drivers, such a breach of duty might be speeding, texting while driving, or driving while intoxicated. For AI developers, the relevant duty will depend on context. For example, plausible precautionary measures that reasonable care would require for preventing misuse of powerful future AI systems might include safeguards to prevent the system from responding to illicit requests, red-teaming to identify and eliminate jailbreaks, structuring system access to prevent the safeguard from being fine-tuned away, and cybersecurity measures to prevent the model weights or other sensitive information from leaking. If courts determine that the duty of reasonable care includes these or similar measures, and the model developer or provider fails to adopt one or more of them, and the plaintiff can prove that adopting one or more of those measures would have prevented their injury, then liability would attach to the developer or provider of the model.

For products liability, the test is a bit different. First, liability would only attach to the developer or provider if the harm is downstream of a commercial sale of a product. This will require a threshold determination regarding whether a particular AI system should be classified as products or services. According to the Restatement (Third) of Torts: Products Liability §19(a) (1998), a product is tangible personal property distributed commercially for use or consumption. Mass-produced systems embedded in physical artifacts, like AVs, are clearly products, but the law is less clear for non-embodied software agents. Even if the system is a product, products

liability would not apply in cases where the system user did not receive access to the product as part of the stream of commerce. Open-weights models, for instance, would not be subject to products liability. Nor would liability attach if a system is subject to a cybersecurity breach, after which the cyberattacker uses the system to do harm. Even systems that are made available for free API access would be exempt from products liability.

Second, products liability requires the plaintiff to prove that the product was defective and that correcting that defect would have prevented their injury. There are three types of product defects: design defects, manufacturing defects, and warning defects. The Restatement (Third) of Torts: Products Liability, which summarizes the general thrust of products liability law across multiple U.S. states, defines these as follows.

A product:

contains a manufacturing defect when the product departs from its intended design even though all possible care was exercised in the preparation and marketing of the product;

is defective in design when the foreseeable risks of harm posed by the product could have been reduced or avoided by the adoption of a reasonable alternative design by the seller or other distributor, or a predecessor in the commercial chain of distribution, and the omission of the alternative design renders the product not reasonably safe;

is defective because of inadequate instructions or warnings when the foreseeable risks of harm posed by the product could have been reduced or avoided by the provision of reasonable instructions or warnings by the seller or other distributor, or a predecessor in the commercial chain of distribution, and the omission of the instructions or warnings renders the product not reasonably safe.

Manufacturing defects are the closest that products liability comes to true strict liability. No matter how careful the manufacturer is in maintaining a quality control process, they can be held liable if an individual unit that comes off the production line deviates from its design specifications, and that deviation causes an injury. But manufacturing defects are a poor fit for software liability, where an individual copy of the software deviating from its design is not a common source of product failure.

The tests of design and warning defects are much more negligence-like. Most states apply a risk-benefit test in assessing the reasonableness of alternative designs. According to this test, an AV would only be defective if a feasible alternative design would have made the product substantially safer without making comparable sacrifices in terms of price, performance, or other relevant product features. A minority of states follow the consumer expectations test, which holds that "a product is defective in design or formulation when it is more dangerous than an ordinary consumer would expect when used in an intended or reasonably foreseeable manner." But

even this approach is qualified with the proviso that some products are "unavoidably unsafe."

> *There are some products which, in the present state of human knowledge, are quite incapable of being made safe for their intended and ordinary use... The seller of such products, again with the qualification that they are properly prepared and marketed, and proper warning is given where the situation calls for it, is not to be held to strict liability for unfortunate consequences attending their rue, merely because he has undertaken to supply the public with an apparently useful and desirable product, attended with a known but apparently reasonable risk.*

This gloss on the consumer expectations test is typically understood to retain elements of the sort of risk-utility balancing that is characteristic of negligence analysis. The reasonableness assessment for information defects involves a similar balancing of the costs and risks of potential instructions or warnings that could have been provided. The difference between design/warning defects and ordinary negligence is that the plaintiff need not prove anything about the conduct of the people responsible for any defect in the product, only that the resulting product was defective according to the respective reasonableness tests.

Case Study: Autonomous Vehicles

Now consider the application of these liability regimes to AVs. In Levels 0 to 2, the vehicle provides respectively "momentary assistance or interventions" (collision warnings, lane departure warnings, and automatic emergency braking), "continuous assistance" (adaptive cruise control and lane-keeping assistance) and "continuous assistance with both acceleration/braking and steering." Nonetheless, for all three of these levels, the driver must remain "fully engaged and attentive." In Level 3, the vehicle "performs all driving tasks," and the driver must "remain[] available to take over any or all tasks if prompted." In Levels 4 and 5, which we may describe as "fully autonomous vehicles," the vehicle is "fully responsible for all driving tasks" and "any occupants act as passengers and do not need to be engaged." That is, at the higher levels of driving automation, drivers become passengers.

While automation can reduce accidents, it is not accident-proof. For example, In March 2018, a self-driving car operated by Uber struck and killed a pedestrian. The vehicle was operating in self-drive mode with a

human safety backup driver in the driving seat. The incident took place at night when the pedestrian was crossing a four-lane roadway with their bicycle, outside of a sanctioned crosswalk. The driver-facing camera showed that the Uber test driver was watching videos immediately before the crash.

Historically, auto liability and insurance have focused primarily on drivers' negligence as the cause of most incidents, with products liability for defects a distant second. In a driver negligence case, the plaintiff must prove that the defendant driver failed to exercise reasonable care and that this failure caused the plaintiff's injury. For example, if the plaintiff can prove the defendant was speeding, drunk, or texting while driving, and that this unreasonable conduct caused the collision, then the plaintiff will generally be able to recover. However, many traffic collisions are what is known as "unavoidable accidents." Unavoidable accidents are not literally impossible to avoid. Instead, they are accidents that would not have been prevented by the exercise of reasonable care. The existence of unavoidable accidents is an inevitable consequence of the reasonableness inquiry built into negligence law. Since a reasonable person would not accept arbitrarily large costs for incremental gains in collision risk, some collisions must occur between drivers (and other road users, like bicyclists and pedestrians) that are both exercising reasonable care. To ensure ability to pay in cases where a driver is liable for another road user's injuries, every U.S. state requires drivers to carry a specified minimum level of coverage for third-party injuries. First-party injury insurance is not required, but many driver's choose to take out first-party insurance to indemnify them in cases of unavoidable accidents and injuries for which they are at fault.

Even prior to the introduction of AVs, automakers could also sometimes be held liable in cases of vehicle collisions. The three liability regimes applicable to automakers are products liability, negligence, and breach of warranty. Products liability can apply in two sorts of cases. First, the collision itself may have been caused by a product defect, like faulty brakes. Second, the injuries sustained in a crash may have been exacerbated due to a product defect, like a malfunctioning airbag. The latter sort of defect is typically analyzed in terms of a vehicle's "crashworthiness."

Breach of warranty claims are somewhat narrower. A warranty is an express or implied promise made as part of the contract of sale. Since they are claims brought under contract law, rather than tort law, they can only be brought by the purchaser of the vehicle against the seller. An express warranty is one clearly stated, such as the description of the product (UCC § 2-313). The key question is whether the express warranty formed part of the "basis of the bargain." An implied warranty is one that exists unless expressly excluded or modified. An implied warranty of merchantability arises by operation of law and states that the product is fit for the ordinary

purposes of its originally intended use (UCC § 2-314). An implied warranty of fitness arises when a buyer requests a product for a particular purpose and is supplied with one; the implied warranty is that the product is fit for that purpose. For this claim, there need not be a defect, but rather proof that the product failed to meet the plaintiff's specifications.

In short, while negligence concerns the defendant's conduct, products liability concerns the defendant's product. While auto manufactures can also be sued for negligence, it is typically much easier to bring a products liability claim because such claims do not require proving that the design or manufacturing process was unreasonable, only that the product was defective.

As more autonomous vehicles are on the road and levels of automation increase, standards of negligence focused on driver conduct and fault may become outdated. Instead, liability regimes may largely "shift from the error caused by human miscalculation or inattention to the design of the automated system." In this respect, it is helpful to distinguish between semi-autonomous vehicles (Levels 0 to 3) from fully autonomous vehicles (Levels 4 and 5). The relative importance of product liability and negligence for autonomous vehicles may change over time, as autonomous vehicles move from Levels 2-3 of automation (semi-autonomous) to Levels 4-5 (fully-autonomous).

In semi-autonomous vehicles, the driver is still in control and expected to take over when necessary. This leaves room for driver negligence.[8] In cases of crashes involving semi-autonomous vehicles, the driver's fault in causing the accident is likely to be taken into account. Therefore, in semi-autonomous vehicles drivers' negligence and manufacturers' product liability are both relevant theories of liability. Since it may be complicated to establish who between the driver or and the semi-autonomous vehicle manufacturer is responsible for the accident, and some drivers may be judgment-proof, it is likely that plaintiffs sue both the driver and the

[8] Manufacturer's negligence can also be relevant, for instance in case a manufacturer overlooks an important aspect in designing, manufacturing, labeling, advertising, inspecting or repairing a semi-autonomous or fully-autonomous vehicle. According to Oberly, a manufacturer's duty of reasonable care concerns "the design of their automobiles to avoid unreasonable risk of injury ... to minimize the severity of injury in the event of an accident ... to construct their vehicles without latent or hidden defects." See also *Molander v. Tesla*; amended complaint, where claimant claimed that "Tesla had a duty to use reasonable due care in the design, manufacture, assembly, packaging, testing, fabricating, analysis, inspection, merchandising, marketing, distributing, labeling, advertising, promotion, sale, supply, lease, rental, warning, selection, inspection, and repair of the 2019 Tesla Model 3." Another example of breach of duty is if a manufacturer only tests a braking system on dry roads, or does not act promptly after discovering a potentially dangerous software problem.

manufacturer. That is, "partial autonomous systems will shift some, but not all, of the responsibility for accident avoidance from the driver to the vehicle."

Negligence is assessed based on the reasonableness of the conduct of the driver. There is substantial disagreement regarding the standard of care for the operation of an autonomous vehicle. Some analysts point out that "[i]n crashes that involve drivers reasonably relying on a car's ability to control itself, there may not be an at-fault driver for the victim to sue." More generally, one scholar has proposed that "courts . . . focus on the ability of the person to prevent the accident, rather than what the driver was doing prior to the accident—otherwise the utility of these vehicles could be greatly diminished." Conversely, industry representatives have argued that for semi-autonomous vehicles "complete reliance on such prophylactic safety devices is likely to be seen as unreasonable." These arguments imply that "the blame falls on the driver in accidents occurring when" such semi-autonomous driving features "are activated." Similarly, it has been observed that "where vehicles are not operating in autonomous mode, but are being driven by a human"—such as in the case of Level 2 and Level 3—"the driver will ordinarily still be subject to liability even in the context of an autonomous vehicle accident." This conclusion also follows from a precedent, Brouse v. U.S, involving a collision between two planes—one operating in autonomous mode—where the court decided that "[t]he obligation of those in charge of a plane under robot control to keep a proper and constant lookout is unavoidable."

The considerations about the driver's fault in semi-autonomous vehicles have a second consequence. They open the door to defenses of contributory or comparative negligence and assumption of risk—and to the argument that the driver's conduct was unforeseeable misuse—when the driver sues the manufacturer on a product liability theory. The proximate cause or scope of liability element common to both negligence and products liability claims would not be satisfied if the injuries sustained by the plaintiff were not a foreseeable consequence of the defendant's conduct or the defect in the defendant's product. There is disagreement over what might constitute misuse in the context of autonomous vehicles. According to the law firm Jones Day, examples of misuse defenses include "disregard of explicit training and warnings" and "failure to accept an update." By contrast, one scholar argued that a driver is not misusing an autonomous vehicle "simply by doing other activities while behind the wheel," and that this defense should be reserved to more severe instances, such as if the driver modifies the vehicle that causes the technology to malfunction.

Contributory/comparative negligence is an affirmative defense, which requires the defendant to prove that the plaintiff breached their duty of care,

and that breach caused their injury. The traditional common law rule—contributor negligence—was that plaintiff negligence was a complete bar to recovery. Most states have now adopted some form of comparative negligence, which generally reduces the defendant's liability in proportion to plaintiff's share of responsibility, using a procedure known as fault allocation. This fault allocation only applies to what are known as individualism injuries, where both the plaintiff's negligence and the defendant's tortious conduct were but-for causes of the entire injury.[9] If a driver's injuries in a collision are exacerbated by the fact that they are not wearing a seatbelt, incremental injuries resulting from the lack of a seatbelt would be divisible from those that would have happened anyway. Even plaintiffs in strict contributory negligence jurisdictions would be allowed to recover for any injuries they would have suffered even if they had been wearing a seatbelt, provided they otherwise exercised reasonable care.

Assumption of the risk is a subtler doctrine that provides more protection to defendants. The broad concept is that the plaintiff either expressly assumed the risk as part of an explicit contractual agreement or that assumption of the risk can be inferred from participation in an inherently risky activity. Doctrinally, assumption of the risk can play out in two distinct ways, depending on the jurisdiction. In jurisdictions that retain primary assumption of the risk, the rule is that defendant's actually have no duty to protect plaintiff's from risks that they voluntarily assumed. This means that the plaintiff cannot satisfy the prima facie case for liability and so cannot recover anything. Other jurisdictions treat assumption of the risk as a form of plaintiff fault, analyzing it under comparative negligence principles. This is known as secondary assumption of the risk.

In fully-autonomous vehicles, the human "passenger" would typically not be a factor in the liability determination, as the human loses all control over the operation of the vehicle. This can correspond to a shift in liability from driver's negligence to manufacturers' strict liability. While driver negligence can still be central in Level 2 and Level 3 vehicles, it is likely to lose importance for Level 4 and Level 5 vehicles.

More specifically, there is debate over what negligence would mean in the context of fully-autonomous vehicles. Some scholars argue that negligence cannot be applied to fully-autonomous vehicles, as they "lack[] direct human input … [and] can't be compared to the reasonable person." Traditional interpretation of the elements of negligence "may need to be revisited" for autonomous vehicles since they "will likely be different from the status quo." In sum, assigning responsibility to the owner of a fully-

[9] A but-for cause is an action or event without which the outcome in question would not have occurred.

autonomous vehicle is "problematic," unless at the time of purchase they agreed to assume the risk.

In contrast, other scholars have suggested theories that would entail the application of negligence theories to fully-autonomous vehicles. According to some scholars, California interprets the notion of "driver" very widely, so much that someone could be considered "driving" without having "actual physical control." It thus seems possible that "passengers" in Levels 4-5 might be considered "drivers" for liability purposes under current frameworks. Other scholars argue that the autonomous vehicle itself may qualify as a "driver." In a letter in response to Google's request for clarification, NHTSA clarified that "Because Google's SDV [self-driving vehicles] design purposely does not have any mechanism by which human occupants could steer or otherwise "drive" the vehicle, it would be difficult in several instances to determine who the "driver" would be in its SDV . . . . NHTSA will interpret "driver" in the context of Google's described motor vehicle design as referring to the SDS [Self-Driving System], and not to any of the vehicle occupants." Similarly, Lemley & Casey argue for the importance of evaluating the safety of autonomous vehicles by the same standard that we apply to human drivers and therefore suggest applying comparative negligence to manufacturers of fully-autonomous vehicles. Anderson & Brown similarly suggest introducing a manufacturer negligence standard, under which the vehicle's would be assessed according to a "reasonable human driver" test.

More likely, driver's negligence will lose relevance in the context of fully autonomous vehicles in favor of products liability. Among the possible defects, legal scholars point to design defects as the one most likely to be invoked with respect to autonomous vehicles.[10] As one scholar observed, "if someone wanted to bring a lawsuit against a manufacturer for how an autonomous vehicle was programmed, they would likely assert a design defect." While it may be difficult for plaintiffs to prove design defects, it is still likely to be the most viable pathway to manufacturer liability. For instance, plaintiffs may be able to bring product liability claims based on the existence of a design defect when in the case at hand there is a late or otherwise inadequate take-over warning in the semi-autonomous vehicle, or a flaw in the original design of the software installed on the semi-autonomous or fully-autonomous vehicle that results in a collision.

As discussed above, a product is considered "defective in design when the foreseeable risks of harm posed by the product could have been

[10] For instance, in litigation involving Tesla's Autopilot (corresponding to Level 2 of automation), plaintiff claimed that the vehicle "was defective because its design was a substantial factor in causing ... injuries ..., and because it did not perform as safely as an ordinary consumer would have expected it to perform when used or misused in an intended or reasonably foreseeable way."

reduced or avoided by the adoption of a reasonable alternative design." Recall that this is evaluated according to two tests: the risk–utility test and the consumer expectation test. Many questions remain open about how each of these tests will apply to AVs. With respect to the consumer expectations test, a jury recently concluded that "the Autopilot" (Tesla's Level 2 semi–autonomous technology) "is one about which an ordinary consumer can form a reasonable safety expectation." Nonetheless, it still remains unclear to what extent a consumer can reasonably expect that a semi–autonomous vehicle drives itself. A scholar observed that, while it could be claimed that a semi–autonomous vehicle was designed to detect objects and obstructions around the car, plaintiffs are unlikely to prevail. Semi–autonomous vehicles were not designed to be a complete substitute to human intervention and therefore consumers are not able to reasonably expect them to detect large moving objects in all circumstances. For the risk–utility test, it remains an open question to what extent more advanced autonomous driving technologies (e.g., LiDAR) could show the mechanical feasibility of a safer alternative design (e.g., as compared to Autopilot).

To date, no court has found manufacturing defects in software. It is not even clear what would count as a manufacturing defect in a pure software product, since the nature of information goods is that they can be perfectly copied.[11] It is conceivable, however, that plaintiffs harmed in AV collisions may be able to bring claims for manufacturing defects if they suffer harm due to physical defects in the embodied systems, such as in the sensors or in the takeover spy. More speculatively, a manufacturing defects theory may be available in cases an incorrect version of the operating software is installed on the semi–autonomous or fully–autonomous vehicle.

In addition to products liability, manufacturers are also exposed to breach of warranty claims. For instance, if a manufacturer advertises a vehicle as fully–autonomous when it only has limited semi–autonomous driving capabilities, or describes the system in online marketing as if it were able to automate tasks that it cannot actually undertake, this situation may be construed to establish an express warranty. While autonomous vehicle manufacturers could "provide buyers with contractual limited warranties and disclaim all other warranties," some scholars recommend bringing warranty claims "when manufacturers overstate their autonomous capabilities." Conversely, other scholars have maintained that "unless the manufacturer had promised an accident–proof vehicle, the choice made by the algorithm could not be connected to an "affirmation of fact or promise" from the manufacturer giving rise to an express warranty."

[11] Copying errors are possible, but are not anticipated to be a major source contributor to AV collisions.

Alternatively, in a case where the autonomous vehicle has a latent defect that compromises its correct functioning, a plaintiff could claim a breach of an implied warranty of merchantability, as the plaintiff did in Hsu v. Tesla. Finally, if a buyer requests an autonomous vehicle with specific capabilities (e.g., operates in all areas or in certain conditions), this situation could lead to breach of implied warranty of fitness if the vehicle does not in fact have those capabilities.

There is debate over how important and successful breach of warranties claims might be in the context of autonomous vehicles. Some analysts predict that claims based on warranty theories of liability are likely to increase. Conversely, and more persuasively, other analysts argue that warranties are unlikely to be particularly important to the development of autonomous vehicles because—except for the implied warranty of merchantability—they can generally be disclaimed, and the implied warranty of merchantability has merged into strict liability in most jurisdictions.

As we move towards an automated future, it is then likely that manufacturers—rather than drivers or users—will be increasingly exposed to claims. For instance, in the case of autonomous vehicles, both the driver of the autonomous vehicle and the driver of a non-autonomous vehicle or non-motorist injured in an accident could bring a products liability claim. Potential defendants include the vehicle or component manufacturers, distributors, suppliers, retailers, and anyone else in the chain of distribution, such as hardware vendor, software licensor, mobile network operator. Under the formalism of products liability, plaintiffs could elect to sue any commercial seller in this chain and will likely elect to sue the party with the deepest pockets. If legislation were to establish liability insurance requirement for particular players in the AV distribution chain, plaintiffs would have strong incentives to sue them. In any case, the costs of insurance or, alternatively, the expected costs of liability, are likely to fall on a range of commercial entities, as well as the end customer, with the precise incidence depending on the elasticities of supply and demand in the relevant market.

There have also been proposals for specific legislative or doctrinal reforms to accommodate the changing risk landscape presented by AVs. For example, Abraham & Rabin suggest that, once Level 4 and 5 vehicles are widely adopted (constituting twenty-five percent of all registered vehicles), a "Manufacturer Enterprise Responsibility" should apply. Their proposal has two parts. First, the manufacturer is liable for all bodily harms, except in cases of substantial comparative negligence. Second, AV owners would continue to purchase conventional auto insurance to cover damage to property and their own losses due to theft. Vladeck also endorses strict liability for AV manufactures, even if they prove substantially safer than

human driver. Wansley proposes holding autonomous vehicle companies liable for all crashes, regardless of fault, cause, or comparative negligence. Their rationale is that the companies would then internalize the costs of preventable crashes and thus be incentivized to make all cost-justified investments in safety. Some scholars even suggested assigning personhood to autonomous vehicles, which "should, like a corporation, be considered a legal person, with the same rights and duties as a human being."

Other proposals address the concern that automotive vehicle companies will incur substantial liability for defects, even as those defects cause fewer accidents, fatalities, and injuries than a human driver would. For example, Geistfeld proposes that, during the transition to increasingly autonomous vehicles, auto manufacturers should be insulated from liability for design defects if premarket testing shows that the vehicle performs at least twice as safely as conventional vehicles and consumers are warned of residual risks. Anderson & Brown propose reforms that aim to be "less costly to both victims and manufacturers. Their proposal would rely on "(1) a manufacturer liability standard that assesses the vehicle's actions under a "reasonable human driver" standard or, in the alternative (2) a victim compensation fund that allows those injured to bypass courts and product liability entirely. While any specific proposal is unlikely to be implemented, market participants should not discount the possibility of substantial reforms to existing liability rules, given the transformative nature of fully autonomous vehicles.

Disruptive legislation or regulation may also concern insurance. Scholars have advanced some proposals regarding insurance of autonomous vehicles. Some have called for a "federal regulation of state-level insurance" and observed that "shift in responsibility from the driver to the manufacturer may make no-fault automobile-insurance regimes more attractive", in some cases supporting "a modified no-fault insurance system, which would treat a fully autonomous vehicle's manufacturer the same as a pure no-fault jurisdiction would treat an at-fault driver when the fully-autonomous vehicle's malfunctioning technology causes an accident."[12] Finally, there is some support for requiring insurance at the point of purchase for fully-autonomous vehicles.

Note that the ramp up in AVs is likely to coincide with the phaseout of internal combustion engines, leading to a decline in revenue from gas taxes

12 No-fault insurance is the alternative to a fault-based liability scheme and "removes the challenge of searching for fault in a fully autonomous landscape." The consequence is that a policyholder's damages are paid regardless of who was ultimately responsible for the accident. The downside of no-fault insurance is that it dampens the incentives of the relevance actors to invest resources in efforts to reduce the likelihood and severity of injuries.

and thereby to a further decline in funding for road and infrastructure repair. This decline in revenue might be addressed by taxing auto manufacturers directly for road damage—such a tax may become politically plausible via the increased responsibility, road monitoring data and revenue generated for auto manufacturers by AVs.

Implications for Auto Insurance

Increasing automation and the relevant shift in liability are likely to have an impact on the insurance market. For instance, for Levels 2 and 3 of driving automation, personal auto insurance will probably continue, in the near-term, to play an important role in addressing cases of driver's negligence with semi-autonomous vehicles. Levels 2 and 3 seem to "require some degree of hybrid coverage for product liability issues, as well as traditional tort-based coverage for operator error or negligence issues."

Levels 4 and 5 are likely to be treated differently in a few ways. First, automation is likely to greatly curtail demand for individual auto insurance. As Level 4 and Level 5 AVs increase their market penetration, demand for private auto insurance is likely to diminish, as drivers will not be the primary bearers of liability. Some scholars even argue that, "[i]f these technologies reduce crashes sufficiently, it is possible that the very need for specialized automobile insurance may disappear entirely. Injuries that result from automobile crashes might be covered by health insurance and homeowner's liability insurance, in the way that bicycle crashes or other crashes are now covered." Conversely, some insurance industry analysts have argued that individual auto-owner insurance will remain the best solution for AVs. Nonetheless, even if personal auto insurance claims will be dominant, secondary subrogation claims against manufacturers for product defects are likely.

Second, as autonomous vehicles are likely to increase the liability exposure of manufacturers, product liability insurance for manufacturers may increase in importance, when compared to personal auto insurance. In Levels 4 and 5, the true "driver"—the party actually controlling the vehicle—is not the human but the vehicle itself. Therefore, there is less reason to underwrite insurance or impose liability on this basis of human driver error. As a result, insurance coverage may shift from drivers to the automakers and software companies responsible for the development and maintenance of various autonomous-driving technologies. Lior predicts that "the main burden of purchasing insurance policies should be put, at least initially, on the company side of the AI transaction" because "insurers will be less able to regulate the driver's behaviour, nor will they be able to mitigate risks of moral hazards."

However, it is worth noting that AV manufacturers will be much better positioned to self-insure against the risk of liability across their entire vehicle fleet than individual drivers are for their idiosyncratic risk of being responsible for a collision. The rationale for legal requirements to acquire liability insurance will also be substantially weaker when deep-pocketed corporations are the potentially liable party rather than individual drivers, many of whom would otherwise lack the ability to pay out even modest damage awards. This suggests a diminished overall role for auto collision liability insurance policies, with the demand for manufacturer liability coverage failing to compensate for the fall in driver insurance.

Third, driving automation will likely put downward pressure on insurance premiums. Analysts have pointed out that "by reducing the risk of human error, autonomous vehicle technologies can reduce the incidence of crashes. This will, in turn, reduce automobile insurance costs." Similarly, "because there will be fewer accidents, there will be lower medical and crash damage repair expenditures by the insurance companies, eventually leading to reduced motor vehicle insurance premiums." However, "we cannot count on a linear reduction in premiums" and "it may take years before the presence of autonomous vehicles affects premium rates to any noticeable degree." In contrast, some scholars argue that premiums for traditional vehicles will rise. As autonomous vehicles reduce the number of accidents, insurance policies for traditional vehicles may become more expensive. "Once … insurance companies see a huge drop in claims because of autonomous cars, the insurers may charge car owners far more to operate traditional (as opposed to driverless) cars and that will create a huge consumer push for driverless cars."

Fourth, driving automation may induce changes in underwriting criteria. "[M]any of the traditional underwriting criteria, such as the number and kind of accidents an applicant has had, the miles he or she expects to drive and where the car is garaged, will still apply, but the make, model and style of car may assume a greater importance." Also, it may alter the information landscape for insurers. "Rather than relying on a driver's statements, insurance companies may begin to more heavily weigh information provided by electronic control modules in vehicles, otherwise known as 'black boxes.'" Such a trend is already visible today in the rise of so-called "telematics insurance."

In sum, the widespread deployment of Levels 4 and 5 AVs is likely to substantially shrink the market for auto liability and collision insurance. This is for two primary reasons. First, liability will largely shift from drivers, who are currently required to carry insurance, to manufacturers, who are better positioned to self-insure. Second, as AV systems improve, a reduction in the volume and severity of auto crashes will decrease both demand for collision insurance and the premiums that insurers are able to command. To thrive

in a world of increased automation, insurers will need to seek out new growth markets.

New Risk Exposures and Opportunities for Insurance

Not only does AI have the potential to cause a shift in liability paradigms, but it can also create new risk exposures. This section describes two: AI developers' and AI users' liability for AI agents, and AI systems as a target of AI-powered cyber attacks.

AI Agents and Liability

Recall that four forms of liability are potentially applicable to harms caused by AI agents: human negligence, products liability, strict liability for abnormally dangerous activities, and vicarious liability for torts committed by AI systems.

The application of the negligence and products liability regimes to AI agents is broadly similar to their application in the context of AVs. For negligence, the plaintiff would be required to prove that the human who developed or deployed the agent failed to exercise reasonable care and that this failure caused their injury. Importantly, the scope of negligence inquiry is typically quite narrow. Courts are unlikely to conclude that deploying an AI agent for a particular task is unreasonable, just as courts do not subject every human decision to take an SUV out for a ride to risk-utility analysis. Given that preventing harms from AI capabilities failures, misalignment, and misuse remain largely unsolved technical problems, it may be difficult for plaintiffs to prove that some precautionary measure that a reasonable AI developer or provider would have taken would have prevented their injury. That said, some AI developers and providers will likely fail to adopt industry best practices for mitigating the risks of harms by AI agents and will thereby expose themselves to negligence liability.

Legally, it should be more straightforward to hold malicious users of AI systems liable, as their negligent or intentional misconduct will generally be a direct cause of the plaintiff's injury. Unfortunately, those users will often be judgment-proof. That is, they will often lack the resources to pay out large damages and may also be criminals, terrorist groups, or foreign governments against whom it would be difficult to enforce a damages award. Enforcing requirements to carry liability insurance on potential misusers may also prove difficult, especially for users of open-weights systems.

Products liability is also likely to be challenging for AI agents. For purely software agents, manufacturing defects are likely to be off the table. For embodied systems, manufacturing defect liability may be viable under

some circumstances, but is unlikely to capture the core novel risks posed by AI agents, particularly those involving misalignment or misuse. As with autonomous vehicles, design and warning defects will be available theories, but are evaluated under negligence-like reasonableness standards. Moreover, products liability only applies to commercial sellers of products. If AI agents are structured as service providers, they may largely escape the products liability regime. For these reasons, products liability is likely to leave many gaps in its coverage of the risks generated by AI agents. Nonetheless, it is likely to play a substantial role in supplementing negligence liability in cases where it is easier to prove that the AI agent is defective than it is to prove that a human failed to exercise reasonable care at some point in the process of developing and deploying the AI agent.

The other two pathways to liability for AI agents, abnormally dangerous activities strict liability and vicarious liability, are more speculative. The abnormally dangerous activities doctrine applies strict liability—that is, liability without fault—to the foreseeable harms arising from certain inherently dangerous activities. While each U.S. state has its own list of recognized abnormally dangerous activities, common examples include blasting with dynamite and other high energy activities, hazardous waste disposal, and activities like crop dusting that involve poisons. A related category of strict liability covers ownership or possession of wild animals and other animals with known dangerous tendencies. There are two broadly adopted formulations of the abnormally dangerous activities doctrine. According to the Restatement (Second) of Torts:

> *§ 519. General Principle*
>
> *One who carries on an abnormally dangerous activity is subject to liability for harm to the person, land or chattels of another resulting from the activity, although he has exercised the utmost care to prevent the harm.*
>
> *This strict liability is limited to the kind of harm, the possibility of which makes the activity abnormally dangerous.*
>
> *§ 520. Abnormally Dangerous Activities*
>
> *In determining whether an activity is abnormally dangerous, the following factors are to be considered:*
>
> *existence of a high degree of risk of some harm to the person, land or chattels of others;*
>
> *likelihood that the harm that results from it will be great;*

inability to eliminate the risk by the exercise of reasonable care;

extent to which the activity is not a matter of common usage;

inappropriateness of the activity to the place where it is carried on; and

(f) extent to which its value to the community is outweighed by its dangerous attributes.

The Restatement (Third) of Torts simplifies the test for abnormally dangerous activities to: "(1) the activity creates a foreseeable and highly significant risk of physical harm even when reasonable care is exercised by all actors; and (2) the activity is not one of common usage."

Developing and deploying AI agents are clearly not currently activities of common usage. Training the frontier systems on which these agents are built is particularly uncommon, at least with current technology, given the enormous computational resource requirements of these systems. However, it is possible, even likely, that training will become more common with advances in algorithmic efficiency. It is also possible that agentic features will mostly be added in computationally-cheap scaffolding layered on top of computationally-costly base models.

Nonetheless, under the Restatement Third's test, the applicability of the abnormally dangerous activities doctrine is likely to turn on whether training or deploying AI agents creates a foreseeable and highly significant risk of harm even when reasonable care is exercised. This question is likely to be controversial. While there is some empirical evidence and strong theoretical arguments supporting the conclusion that reasonable care may be insufficient to reduce the risk of catastrophic AI misalignment or misuse to below levels that would qualify as "highly significant," recognizing *any* software development project as abnormally dangerous would represent a substantial doctrinal innovation. Note also that, even if an AI system's deployment is initially recognized as being abnormally dangerous, if the system has been deployed for long enough to be proven reliable and for people to adapt, it might cease to qualify as abnormally dangerous.

The Restatement Second's formulation is even less likely to support strict liability for AI agents. In particular, factor (f)—the "extent to which its value to the community is outweighed by its dangerous attributes"—is likely to weigh against strict liability. The creators, providers, and users of AI agents are likely to emphasize the great potential social benefits of AI agents in curing diseases, advancing science, and accelerating economic growth. Factor (e)—the "inappropriateness of the activity to the place where it is carried on"—is also unlikely to support the application of strict liability.

This suggests that most courts are unlikely to extend the abnormally dangerous activities doctrine to AI agents, at least absent some major forcing event. Nonetheless, this is a sufficiently plausible doctrinal move that it would be unwise for developers, providers, and users of AI agents to rule out the possibility that their activities will be subject to strict liability. This outcome is particularly likely if it becomes apparent that AI agents are causing substantial harms that is inadequately addressed by negligence and products liability, given the substantial gaps discussed above.

Vicarious liability applies when one legal person, the principal, is held liable for torts committed by another legal person, the agent. The most prominent form of vicarious liability is the doctrine of *respondeat superior*, under which employers are held liable for the torts of their employees committed within the scope of their employment. This doctrine does not require any employer negligence, or other misconduct in recruiting, screening, training, supervising employees. However, the doctrine does exclude the actions of independent contractors and acts of employees that fall outside their scope of employment. The status of an employee versus an independent contractor is not determined by their label in a contract, but rather by the degree of control that the employer has the right to exercise over the manner and means of the agent's performance of the directed tasks. The scope of employment includes acts arising out of employment (i.e., furthering some employer purpose) and acts undertaken in the course of employment (i.e., personal acts that are incidental to and concurrent with the performance of employment functions). Minor deviations ("detours" is the term of art) from tasks that advance the employers objectives will generally not break the scope of employment, but torts committee as part of major departures from work tasks ("frolics") will fall outside the scope of employment.

While no court has recognized vicarious liability for the actions of AI agents, several legal scholars have argued that it is the most appropriate mechanism for assigning liability for the harms caused by agentic AI systems. These scholars argue that vicarious liability would create incentives for human developers, providers, and users of AI agents to better train and guide their agents to prevent harmful actions. Others caution that vicarious liability should be limited to circumstances where the AI agent "operates autonomously in a mission-critical setting or one that has a high possibility of externalizing the risk of failure on others, such as when it is used in a highly interconnected market or to perform a medical procedure."

Regardless of the normative appeal of vicarious liability, there are several practical and doctrinal barriers. First vicarious liability is inherently dependent on the agent as the primary vessel of liability. Employer liability under *respondeat superior* serves as a backstop, allowing injured parties to recover from the deep-pocketed employer rather than suing an employee

who may be judgment-proof. But the employer can only be held liable if the employee is liable. A plaintiff in a vicarious liability case must prove all the elements of the underlying tort for the agent in addition to establishing that the employer is eligible to be held vicariously liable for those torts. But no court has ever held an AI agent or other software-based system liable for any tort. Doing so would require some sort of theory of AI legal personhood. While this theory need not necessarily endow AI agents with rights and privileges attendant to legal personhood, it would need to impose at least some of the duties traditionally linked to legal personhood.

The Restatement (Third) of Agency, which reflects the legal status quo, specifies that "a computer program is not capable of acting as a principal or an agent as defined by the common law. At present, computer programs are instrumentalities of the persons who use them . . . That a program may malfunction does not create capacity to act as a principal or an agent." Lior argues that the programs that were "at present" when the Restatement was drafted are not comparable to advanced AI agents, and so the wording of the restatement should not be a blocking factor to establish an agency relationship with AI agents. Regardless, treating AI agents as legal persons, at least for the purpose of assigning tort liability, would represent a substantial doctrinal innovation. Like the abnormally dangerous activities pathway, it should be considered a live possibility, but not the default pathway.

Further, recognizing AI systems as agents is only one step along the path to vicarious liability. Following the logic of the *respondeat superior* doctrine, the human principal would have to have the right to control the manner of means by which the AI agent accomplishes its tasks. While humans are likely to retain this right as a matter of legal formalism, a major concern is that human principals will have only limited ability to control some AI agents. This could make respondeat superior a poor doctrinal fit. After all, *respondeat superior* is premised, at least in part, on the idea that employers have some practical capacity to monitor their employees and to enforce constraints on their behaviour. If that practical capacity is lacking for AI agents, then it would be reasonable to question the applicability of vicarious liability. If the concern is that the very act of deploying AI agents generates risks that the deployers should bear, that seems like a better fit for the abnormally dangerous activities doctrine analyzed above.

Even if courts are prepared to recognize AI agents as the equivalent of employees for the purpose of *respondeat superior*, it may not be not clear which legal person should be treated as the employer. If the same legal person (which may be a vertically integrated corporation) designs and deploys the system for its own business purpose, then it would be the principal. But if an AI based model is trained by one entity, fine-tuned and scaffolded to be agentic by another, and sold to retail customers by a third

entity, and deployed by an end user, who is the agent's principal? Scholars have suggested that the identity of that principal may be circumstance dependent. At an early stage, or with an unsophisticated user, courts may be most likely to treat the AI developer as the principal. As AI agents become more pervasive and users become more sophisticated, liability may shift to the end user. For systems that are accessed via a successful cyberattack, the hacker would be the principal, but the developer may be found negligent.

Finally, even once a court has determined that the AI agent committed a tort and that it is the agent of a specific principal, there remains the question of whether the agent's tortious conduct is within the scope of employment (or deployment, as it may be). Harms that arise from capabilities failure or misuse would be within the scope of deployment, but misalignment is a tougher case. If the AI system is pursuing the high-level goals of its developer or user, but merely doing so via means of which the principal would disapprove, this is probably within the scope of deployment. But if the AI agent starts pursuing goals other than those intended by its human principal, and that pursuit is more than an incidental deviation from its pursuit of the principal's goals, then this would likely sever the principal's liability.

In sum, there is a great deal of uncertainty regarding the application of agency law to AI systems. As with application of the abnormally dangerous activities doctrine to AI development and deployment, applying vicarious liability for the actions of an AI agent would represent a doctrinal innovation, and should probably not be considered the default outcome. But developers, providers, and users of AI agents cannot rule out the possibility that they will be held liable for harms caused by systems they develop, control, deploy, or use.

AI Systems as the Target of Cyberattacks

Cyberattacks can be used to gain access or cause harm to an AI system. According to the National Institute of Standards and Technology (NIST), an agency within the U.S. Department of Commerce, "adversaries can deliberately confuse or even 'poison'" AI systems. In this context, the United Nations Institute for Disarmament Research (UNIDIR) defines "cybersecurity risks" as "malicious intentional attacks that can derail how an AI system learns and acts."

Cybersecurity is one of the critical categories of risk included in (i) the voluntary risk management protocols adopted by the three major AI labs; (ii) the voluntary commitments rendered by the major AI labs and other tech companies to the White House; (iii) the G7 Hiroshima Process International

Code of Conduct. In an effort to address the risks posed by cyberattacks on AI systems, both President Biden's Executive Order on AI and the EU AI Act establish cybersecurity requirements for regulated models.

Generative AI models are exposed to a variety of attacks. In particular, AI systems may be vulnerable to confidentiality, integrity and availability attacks during development, maintenance, and deployment. The UK Government's Department for Science, Innovation and Technology has identified twelve potential vulnerabilities in AI systems that are exploitable in the design and development phase, and another eight vulnerabilities that are exploitable during deployment. According to NIST, AI systems are potentially exposed to poisoning attacks during design and training, and potentially exposed to adversarial examples and privacy attacks during deployment.

Confidentiality (or privacy) attacks involve the extraction of hidden information about the model, including data. The attacker's goal is generally to learn about the model's structure and thus be able to manipulate it later. There are three types of confidentiality attack.

In model extraction, attackers try to create a facsimile of the model, constituting a form of theft.

In membership inference, attackers study the inputs and outputs of the system to determine if a data sample was part of the training data, potentially revealing sensitive information within the training data.

In model inversion, attackers try to infer sensitive attributes of the training data by analyzing the model's outputs. This process can lead to recovering private information about individuals recorded within the model's training data.

Integrity attacks are attempts to compromise or derail an AI system, often by manipulating the training dataset and cause the system to be less accurate. UNIDIR notes that these attacks are computationally inexpensive. According to NIST, foundation models are especially susceptible to poisoning. It is common to scrape data from a wide range of public sources. Adversaries can control a subset of the training data and cause targeted failure by poisoning as little as 0.001% of the dataset. Integrity attacks may also involve making subtle changes to the inputs of a system, causing the system to misclassify objects.

Availability attacks attempt to impair the functioning of the model at deployment time, slowing it down or completely stopping it. Availability attacks of critical systems generally rely on ransomware, but there is a wider range of potent techniques for rendering AI systems inoperable.

There is evidence that the cybersecurity practices of the leading AI developers are inadequate. According to reporting in the New York Times, OpenAI has disclosed that, in early 2023 "a hacker gained access to the internal messaging systems of OpenAI, the maker of ChatGPT, and stole

details about the design of the company's A.I. technologies." While this particular hack did not implicate the weights of OpenAI's most valuable models, it may be indicative of broader vulnerabilities.

Defending against cybersecurity attacks from determined, well-resourced actors is difficult, expensive, and cumbersome. The measures necessary to secure against such attacks are incompatible with the start-up ethos of today's leading AI developers and would substantially impede progress. For example, every piece of open-source code that an AI developer wishes to integrate into its system would need to go through an extensive clearance process to check for malware and other potential exploits.

These difficulties help to explain cybersecurity expert Dmitri Alperovitch's claim that "In fact, I divide the entire set of Fortune Global 2000 firms into two categories: those that know they've been compromised and those that don't yet know." This observation has been substantiated by leading national security figures, including former FBI Directors Robert Muller and James Comey and former NSA Director Michael McConnell. More recently, Alperovitch and other experts have suggested that advances in cybersecurity practices have enabled some companies to achieve cyber-resilience.

According to a RAND report, "There is rough agreement among cybersecurity and national security experts on how to protect digital systems and information from less capable actors, but there is a wide diversity of views on what is needed to defend against more-capable actors, such as top cyber-capable nation-states." Unfortunately, according to the same report, "the security of frontier AI model weights cannot be ensured by implementing a small number of 'silver bullet' security measures." Rather, "a comprehensive approach is needed, including significant investment in infrastructure and many different security measures addressing different potential risks." While "there are many opportunities for significantly improving the security of model weights at frontier labs in the short term," the report also warns that "securing model weights against the most capable actors will require significantly more investment over the coming years."

There is a robust debate regarding the impact of open source on the vulnerability of AI systems to cyberattack. On the one hand, open-source development could allow attackers to embed malware within open-source models. As discussed below, open-source models can also be readily fine-tuned and used to support cyberattacks on other systems. But proponents of open source argue that it facilitates spotting and correcting of vulnerabilities. Ahead of a client's use of open-source code, insurers might take on the role of screening that code for vulnerabilities, for a cost, before then insuring that code's deployment.

A new opportunity for insurers may consist in offering tailored cyber risk insurance products for AI developers and providers. Some risks that AI developers may be interested in insuring against include (i) the risk of attackers stealing AI models, training data, or the data that users submit to the model endpoint; (ii) the risk of attackers modifying AI models to produce wrong results in a way that benefits them. In the process of developing cybersecurity insurance products for AI developers, insurers will develop models for measuring and mitigating cybersecurity risk. This suggests that, over time, the insurance industry can help promote better cybersecurity practice.

AI Systems as the Instrumentality of Attacks

Cyber offense warrants particular attention as a potential misuse of advanced AI systems, agents or otherwise. Cyber offense is frequently listed as one of the extreme risks posed by AI, threatening national security, commercial stability and individual safety. AI can increase the accessibility, frequency, and destructiveness of cyberattacks. First, AI can lower the barrier to entry for cyberattacks, thus increasing accessibility of cyberattacks. As the NIST Risk Management Framework describes, AI has the "potential" to "discover or enable new cybersecurity risks through lowering the barriers for offensive capabilities." Google DeepMind's Frontier Safety Framework describes this risk as "cyber enablement" ("increasing text generation, programming, and tool-use capabilities in models, combined with improved understanding of cyber offense strategies, could help amateurs overcome difficult steps in the planning and execution of attacks.").

Second, AI can also increase the "success rate, scale, speed, stealth, and potency" of cyberattacks. According to Munich Re, 2024, cyberattacks will "become increasingly automated and personalized, as well as cheaper and faster to distribute at scale in all languages." Google DeepMind's Frontier Safety Framework, describes this risk as "cyber autonomy" ("the automation of such attacks would significantly lower the costs of doing so, and moreover would enable the execution of such attacks at scale"). AI can facilitate the discovery of critical vulnerabilities in hardware, software or data, thus "increas[ing] the pool of options for threat actor." AI could also enable vulnerability discovery in challenging domains, such as embedded micro-code and firmware, decompiled proprietary binaries in closed source enterprise software, hardware device drivers. AI can also lower the cost to develop polymorphic malware that is able to change its features and thus evade detection.

Furthermore, AI can make it easier to write code to exploit these vulnerabilities, including by developing AI-powered co-pilots. Palo Alto Networks notes that co-pilots are "not yet a fully-realized reality" and classify it as a long-term risk. By contrast, medium-term risks revolve around the sophistication of cyber threats, including reconnaissance purposes and refinement of spear phishing. In the context of AI-powered cyberattacks, Lloyds ranks the evidence of vulnerability discovery as "high" and its potential impact as "very high." Finally, AI makes it possible to run millions of systems at a lower cost and in parallel. Lloyds also ranks the evidence and the potential impact of campaign planning and execution as "very high."

AI-powered cyberattacks can even have catastrophic effects that impact entire societies, including destroying critical infrastructure, such as electric grids and water supply systems. Lloyds estimated that AI could cause a "modest increase in the risk of manageable cyber catastrophes" and make state-sponsored espionage and sabotage more effective. Catastrophes could also arise from a failure in the mechanisms designed by attackers to keep the campaign under control. This risk increases with the portion of critical infrastructure that is on the grid. In particular, Lloyds estimates "an increase in lower-level cyber losses", such as: (i) more errors of judgment (such as spear phishing, executive impersonation, poisoned watering holes), due to targeted and fine-tuned attacks; (ii) higher absolute number of losses, as attacks would reach broader audiences; (iii) more industrial or operational technology attacks. Finally, cyber risk also intersects with other AI-related risks, including disinformation, and manipulation of high-value persons, including through spearfishing attacks on persons in leadership positions.

Implications for Insurance

In contrast to the auto sector, both AI agents and AI systems as both targets and instrumentalities of cyber attacks represent major potential growth markets for insurance in the coming years. The risks associated with AI agents and AI-related cyberattacks include potentially catastrophic harms, which even well-capitalized AI developers and providers are poorly positioned to self-insure against. The case for mandatory liability assurance for these risks is also substantially stronger than it is for AV manufacturers, since the liabilities of individual developers and providers are likely to be subject to much wider variance.

Given the large amount of legal and technological uncertainty around AI agents, insurance products may have an important role to play as the industry develops. Moreover, as the capabilities of AI agents improve, the

risks associated with alignment failures and misuse will also grow. This suggests that liability insurance for AI developers and users of AI agents could be a growth market in the coming years. Given that misaligned or misused AI agents may generate risks that are too large for insurance companies to underwrite, one of this report's authors has even suggested that punitive damages should be available in near miss cases of practically compensable harm that are associated with uninsurable risk. That same author has also proposed liability insurance requirements for the training and deployment of advanced AI systems, especially AI agents. If either of those proposals were to come to pass, insurance companies would likely need to invest in evaluations and other methods for estimating the risks of catastrophic misalignment or misuse of AI agents.

Potential models for AI agent liability insurance include existing insurance products for liability associated with harms caused by children and animals. Given the important differences between AI agents and these precedents, however, insurers should proceed with caution in adapting policies tailored to those contexts. AI agents are still a nascent technology. Early in their deployment, the focus may be on harms arising from capabilities failures. But, as with autonomous vehicles, the volume of harm due to capabilities failures is likely to diminish as the technology matures. Alignment failures and misuse, by contrast, are likely to remain substantial sources of risk for AI agents, which insurance can plan an important and enduring role in managing. Aioi Nissay Dowa Insurance has provided one of the world's first AI agent insurance products, covering risks such as GenAI tools infringing copyright.

In the cyber domain, Lloyds forecasts an increase in lower-level losses, including more errors of judgment (such as spear phishing, executive impersonation, poisoned watering holes), due to targeted and fine-tuned attacks; higher absolute number of losses, as attacks would reach broader audiences; and more industrial or operational technology attacks. Lloyds also expects a "modest increase in the risk of manageable cyber catastrophes" and expresses concern that AI will make state-sponsored espionage and sabotage more effective.

Likewise, Pinsent Masons forecast that "AI tools in cyber attacks" will "create greater exposure for insurers as it would seem to follow that the volume of claims notifications will also increase." A Munich Re study also found that the "global cyber insurance market has reached a size of US$ 14bn in 2023 and is estimated by Munich Re to increase to around US$ 29bn by 2027." An Aon report found that information assets could result in a probable maximum loss of $1.16 billion compared to tangible assets. According to the same report, however, only 19% of information assets are covered by insurance, with self-insurance more widely used at 58%. The primary reasons given for not purchasing a standalone cyber security

insurance policy are: coverage is inadequate based on their exposure (38%), premiums are too expensive (37%) and there are too many exclusions, restrictions, and uninsurable risks (29%). This suggest a substantial opportunity for insurers that are able to craft attractive cyber insurance products.

Swiss Re has also raised alarm bells about the risk of "silent AI," a term, inspired by "silent cyber," used to describe the unintended coverage of AI risks by non-AI policies. They recommend "understanding which risks traditional policies already (silently) cover" first. According to their analysis, cyber insurance already exists and could apply to several possible risks related to AI, such as intellectual property theft, digital asset loss, third-party liability for data breaches, and infringement. Swiss Re suggests that the effectiveness and scalability of cyberattacks that utilize AI may warrant an exclusion and separate endorsement, or a change in premium. Law firm Herbert Smith Freehills reports not having seen AI exclusions appear in traditional policies and recommends that insurers take a stance on whether to price in or exclude that risk in their policies.

Tailored cyber risk insurance products for AI developers and providers may also present new opportunities for insurers. Lloyds suggests that AI developers may be interested in insuring against the risk of attackers stealing AI models, training data, or the data that users submit to the model endpoint in addition to risks associated with hostile actors modifying AI models to produce wrong results in a way that benefits them.

As in other areas, insurance premiums that adapt to risk exposure can incentivize adoption of better security and risk management measures. For example, the Healthcare Cybersecurity Benchmarking Study 2024 found that higher cybersecurity preparedness and resilience—specifically, adoption of the NIST cybersecurity framework—corresponded to lower increases in cybersecurity premiums.

The Societal Benefit of Insuring AI Risks

Insurance serves as a powerful catalyst for societal stability and progress. In addition to its core risk-spreading function, insurance premiums can send salient price signals that encourage responsible innovation practices that balance the benefits of automation against the risks. The availability of insurance to spread risk can also encourage the development, and diffusion, of innovations that might otherwise seem excessively risky.

Consider the case of AVs: the shift in liability from driver's fault to manufacturer's liability may deter manufacturers from developing autonomous driving technologies. Scholars have pointed out that "manufacturers may be reluctant to introduce technology that will increase their liability." These risks could hamper or delay the deployment of technologies with vast social benefits. Insurance of manufacturers and other technology suppliers might help counter this disincentive. After all, scholars have argued that "[t]he insurance industry is the institution best suited to monitor and adapt to evolution in the AI landscape due to its ongoing collection and review of data, as well as its ability to implement change faster than the traditional tort system." One caveat to this analysis is that AV liability risk can be expected to decline over time as the capabilities of the relevant algorithms improve, transportation systems are redesigned to accommodate Avs, and fewer human-driven cars are on the road to cause problems for AVs. This suggests that the social benefits of insurance of AV manufacturers are likely to be concentrated in the early years of widespread AV deployment, when the risks associated with the technology and still relatively large and uncertain.

Similarly, insurance for AI agents can have a positive impact in supporting the development of this technology, by shouldering the risks faced by developers and enabling smaller AI developers to develop agents. According to Lior, "it seems reasonable to assume that applying" strict liability, "would lead to many of the companies with fewer financial resources removing themselves from the market out of fear of bankruptcy." Further, "insurance law has significant value, allowing society to reap the benefits of a strict liability regime without the danger of stifling innovation."

As in other areas, cybersecurity premiums that price risk can incentivize adoption of better security and risk management measures. For example, the Healthcare Cybersecurity Benchmarking Study 2024 found that higher cybersecurity preparedness and resilience—specifically, adoption of the NIST cybersecurity framework—corresponded to lower increases in cybersecurity premiums.

Conclusion

AI-driven automation is likely to pose significant challenges for existing liability regimes and insurance practices. In some domains, like auto collision risk, the likely shifts in both the liability rules (from drivers to vehicle manufacturers) and the frequency and severity of accidents (both down due to advancing AI capabilities) will tend to curtail demand for insurance. While a market for manufacturer liability insurance may emerge as AVs are deployed more widely, that market is likely to peak at a size smaller than that of current driver liability and first-party collision insurance. Then, as AV capabilities continue to improve, that market will only shrink further over time.

However, other domains, particularly AI agents and AI-related cybersecurity threats, are likely to emerge as major growth markets for insurers. Even absent major law or policy changes, the liability risk exposure for developers and providers of AI agents is likely to be both substantial and high-variance, generating substantial demand for new insurance products. Moreover, both AI agents and AI-driven cybersecurity threats are rooted more in alignment failures and misuse than they are in capabilities failures. Whereas capabilities progress is likely to decrease risk in AVs, it will likely increase risks associated with AI agents and cybersecurity. This suggests that insurance demand for AI agents and AI cyber risks are likely to continue to grow as AI systems mature. With sensible policy changes, like expansion of strict liability or liability insurance requirements for AI agents and AI cyber risks, the demand for these new insurance products is likely to be even more robust.

The case is strong for reforms to liability and insurance law to accommodate the new risk landscape. On the one hand, existing product liability regimes might be too strong for AVs. If Level 4 and Level 5 AVs prove safer than human drivers, as early data suggests, then holding manufacturers liable when their systems do fail may, by discouraging the deployment of AVs, actually cause more collisions, injuries, and deaths. Applying a reasonable human driver standard, instead of a reasonable alternative design product defect standard, would level the playing field between human and automated driving and allow both to compete on price, convenience, and safety. Alternatively, it might make sense to level up standards, applying a stricter standard to both human drivers and AVs. Although the latter may be desirable, the resistance to imposing greater liability on human drivers may prove too strong. Meanwhile, lowering the AV standard to match human negligence would likely depress demand for auto collision insurance even further. Likewise, liability insurance requirements may be more difficult to justify, as liability shifts from drivers to automakers,

who are better positioned to demonstrate the ability to pay out a stream of damages awards out of annual revenues, thereby self-insuring.

On the other hand, AI agents and cyber risk are domains where stronger liability rules seem warranted. AI alignment and the prevention of misuse are difficult and unsolved technical and social problems. Merely exercising reasonable care, as defined by the narrowly-scoped standard breach of duty analysis in negligence cases, is unlikely to offer adequate protection against the large and novel risks presented by AI agents and AI-related cyber attacks. Likewise, products liability, even where it applies, is of little use when no one has solved the underlying technical problem, so there is no reasonable alternative design at which to point so as to establish a design defect. These deficiencies point to the need for true strict liability, either via an extension of the abnormally dangerous activities doctrine or holding the human developers, providers, and users of an AI system vicariously liable for their wrongful conduct. These policy changes could be adopted by courts, exercising their common law powers, or via state or federal legislation. In a mirror image of the story for AVs, these changes in liability law would further stimulate demand for insurance products covering liabilities associated with AI agents and AI-related cyber risk.

Moreover, and again contrasting with the case of AVs, a compelling case can be made for new liability insurance mandates for AI agents and AI-related cyber risks. Unlike AV collision risk, the risks in these domains include potential society-wide catastrophes, are likely to exhibit high variance both across products and over time, and are likely to grow rather than diminish as AI capabilities improve. Also, unlike the auto industry, many important players in AI development are venture-funded startups that may lack the ability to pay out large damages awards. All these features point to mandatory insurance as an important tool for both ensuring victim compensation and sending clear price signals to AI developers, providers, and users that promote prudent risk mitigation. Insurance requirements may also raise the salience of liability risk for harms from AI misalignment and misuse that developers and providers might otherwise dismiss or neglect. Unlike the expansion of strict liability, insurance requirements could only be enacted pursuant to new legislation.

Finally, it may be worth considering expanding the availability of punitive damages as a means of regulating uninsurable risks. That is, AI agents or AI-involved cyber attacks may result in harms so large (e.g. the destruction of essential national infrastructure) that it would not be practically feasible to enforce a compensatory damages award, just as is true of war and some acts of terrorism today. Even if liability insurance is required, there will be some limit to the size of risks that insurers are willing and able to underwrite. Harms that exceed this threshold cannot be expected to result in a compensatory damages award. Accordingly, AI

developers and providers may have inadequate incentive to invest in costly measures to mitigate such risks, even if those investments would be expected to produce positive social returns.

One means of addressing this problem would be to allow plaintiffs in "near miss" cases (of practically compensable harm that are associated with uninsurable risk) to recover not just for the harm they suffered, but also for the uninsurable risks that the defendant generated. For example, if the defendant developer's AI agent caused $50,000 worth of damage, but also generated a 0.001% risk of a mass casualty event associated with $5 trillion in damage, then the plaintiff would be able to recover $50 million in punitive damages along with their $50,000 in compensatory damages. Alternatively, a portion of the punitive damages could be diverted to a fund supporting efforts to mitigate uninsurable AI risks. In either case, the function of the punitive damages would be to compel AI developers and providers to account for the full range of risks generated by their systems, including risks of harms that are too large to be practically compensable.

Insurers would be implicated in this liability regime in two ways. First, given that uninsurability is the critical threshold above which punitive damages would be used as the primary risk mitigation mechanism, insurers would have an important role to play in determining the maximum insurable risk. Second, the demand for insurance in a liability regime that included "near miss" punitive damages and liability insurance requirements would be substantial. The insurance industry would need to work hard to quantify the risks associated with powerful AI systems in order to underwrite the full range of liability risks, including indirect "near miss" liability.

To conclude, AI progress presents a watershed moment for the insurance industry. Whilst traditional markets like auto insurance may shrink, new frontiers in AI agent and cybersecurity coverage are poised for explosive growth. The legal landscape is shifting, demanding innovative approaches to liability and risk management. From strict liability regimes to mandatory insurance and creative punitive damages, the toolkit for governing AI risks is expanding. Insurers now face a dual challenge: estimating the difficult-to-quantify risks of advanced AI systems, whilst simultaneously developing products to underwrite those risks. The insurance industry must not simply adapt to AI—it must grow to become a critical pillar of responsible AI governance. The future of AI safety may well hinge less on the developer's code than on the actuary's spreadsheet.

Acknowledgements

The authors offer their deepest thanks to those colleagues who provided valuable suggestions informing this report, including Lindsay Chadwick, Masahiro Indo, Derek Lietz, Phil Norris, Cullen O'Keefe and Bernardo Perez Orozco. We are also very grateful to Aioi R&D Lab Ltd for funding the project.