The AI Liability Puzzle and a Fund-Based Work-Around

Olivia J. Erdélyi
School of Law
University of Canterbury
Christchurch, New Zealand
olivia.erdelyi@canterbury.ac.nz

Gábor Erdélyi
School of Mathematics and Statistics
University of Canterbury
Christchurch, New Zealand
gabor.erdelyi@canterbury.ac.nz

ABSTRACT

Certainty around the regulatory environment is crucial to facilitate responsible AI innovation and its social acceptance. However, the existing legal liability system is inapt to assign responsibility where a potentially harmful conduct and/or the harm itself are unforeseeable, yet some instantiations of AI and/or the harms they may trigger are not foreseeable in the legal sense. The unpredictability of how courts would handle such cases makes the risks involved in the investment and use of AI incalculable, creating an environment that is not conducive to innovation and may deprive society of some benefits AI could provide. To tackle this problem, we propose to draw insights from financial regulatory best-practices and establish a system of AI guarantee schemes. We envisage the system to form part of the broader market-structuring regulatory framework, with the primary function to provide a readily available, clear, and transparent funding mechanism to compensate claims that are either extremely hard or impossible to realize via conventional litigation. We propose at least partial industry-funding, with funding arrangements depending on whether it would pursue other potential policy goals.

KEYWORDS

AI Liability; Guarantee Schemes

ACM Reference Format:
Olivia J. Erdélyi and Gábor Erdélyi. 2020. The AI Liability Puzzle and a Fund-Based Work-Around. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), February 7–8, 2020, New York, NY, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3375627.3375806

1 INTRODUCTION

With proliferating AI-human interactions, questions of AI liability are in the forefront of policy debates. Can a bank using an AI-enabled lending decision-making system that unexpectedly turns out to unlawfully discriminate customers successfully sue the provider of the system? Who is liable if an autonomous vehicle (AV) hits a pedestrian or is involved in a crash? What happens if an AI engages in criminal actions owing to, say, an unexpected value alignment problem of the sort described in Schreier's *Robot and Frank* or the canonical *paperclip maximizer* doomsday scenario?

While each of these questions relate to different domains of legal liability—contractual, tort, and criminal liability, respectively—their core inquiry is the same: Who should be held accountable if something goes wrong with an AI and based on what rules? Legal literature offers conflicting accounts on how best to go about AI liability and the legal system's overall ability to adapt.

[16] synthesizes the literature on selected aspects of civil and criminal liability. [13] maintains that the existing US contractual and tort liability system strikes the right balance between ensuring safety and incentivizing innovation, so it can handle the liability of *sophisticated robots* (those having some degree of connectivity, autonomy, and maybe machine-learning (ML) ability).

Regarding AVs, [22] stresses the importance of clarity and predictability of liability regimes to facilitate the assessment of risk exposure. Others argue for subjecting AVs or AI more broadly to strict liability—commonly some type of products liability [10, 22]. [34] advocates a strict liability regime detached from notions of fault to overcome situations where fault is impossible to establish, and—like [14] for autonomous robots (i.e., those with an ML component)—touches upon the issue of legal personality. [11] is concerned about ludicrous expenses involved with complex products liability suits, pre-trial settlements, product recalls, and punitive damages, pressing for a meticulous application of the negligence doctrine.

[15] doubts whether any of the classic US tort doctrines—negligence and the various forms of strict liability—is up to allocating liability for wrongdoings of truly autonomous robots. This is because foreseeability is a central element of each, but due to complex non-linear interactions between intricate robots and their convoluted, unpredictable environment, neither robots' actions nor the potential harms they may cause are foreseeable in the sense required by law. Regarding AVs and autonomous robots and mostly in the context of US tort law, other commentators voice similar concerns about potential liability gaps and the implications of the resulting uncertainty surrounding the legal liability of AI systems [2, 4]. This unpredictability of foreseeability makes it even harder to evaluate the chances of success of litigation and hence exposure to liability, adding to the uncertainties that flow from the inconsistency of jurisprudence during the typically significant time lag needed for the legal system to adapt to novel technologies. The resulting problems—known in law and technology literature [27]—are inhibition of innovation and adoption of new technologies, in extreme cases reaching as far as shutting down entire emerging markets.

We would like to restrict the focus of this AI liability debate to the analysis of the foreseeability concept's ability to serve as a means to limit and attribute legal liability to AI systems, highlighting a potential conceptual problem. We argue that this problem is

AIES '20, February 7–8, 2020, New York, NY, USA

ACM ISBN 978-1-4503-7110-0/20/02…$15.00
https://doi.org/10.1145/3375627.3375806

common to all types of legal liability, currently most relevant to certain types of embodied or disembodied ML-based AI systems, but understand the term *AI system* broadly, including any present and future AI technologies that may pose similar challenges. We also move the discussion beyond US tort law, embodied AI systems, or particular AI applications—indeed beyond any national analysis and law in general, for the following reasons:

AI is just the most recent among a series of technological innovations, which have fundamentally impacted our societies and economies over time. Yet due to its rapid pace of development, massively transformative nature, and other changes our world has undergone—most notably globalization—AI is anticipated to affect humanity and our environment more intensely. Recognizing this, major national AI strategies and international policy initiatives aim to forge an innovation-friendly, enabling regulatory environment, capturing benefits and minimizing potential risks AI may bring [1, 7, 9, 25]. They converge on the point that successful societal adoption of AI requires trust on the part of society. Trust hinges on at least some level of certainty about how AI will impact society and the economy: developers need to be able to assess the risks inherent in bringing a new product on the market, while consumers and other users must be assured that their use is reasonably safe. Without trust and certainty, there can be no market for emerging technologies. Certainty itself flows from a safe, transparent, and flexible regulatory environment that supports innovation, but designing one is neither a purely legal nor an exclusively national enterprise.

From an economic perspective, regulatory frameworks structuring our economies and market imperfections crucially determine society's benefits from technological innovation [18, 32]. In our reality of imperfect markets, technological innovation is not necessarily Pareto-improving. In the absence of redistributive measures, it can actually aggravate inequality and decrease overall welfare. Inequality-related problems can only be effectively tackled by a holistic approach involving a complete and systematic revamp of market-structuring regulatory frameworks, which legal liability regimes are part of.

The above cited international AI policy documents, and—based on a review of international relations literature—[6] also underscore the necessity of international coordination and cooperation in the AI domain. The core of the arguments here is that issue areas with transnational impact—such as AI—are impossible to effectively regulate by isolated national measures. This is because their inevitable fragmentation and divergence invoke inefficiencies and tensions in international policymaking, negatively affecting domestic regimes and shattering both national and international actors' faith in such approaches.

These arguments call for a holistic, multidisciplinary, and transnational perspective, forbidding an isolated legal or nationally focused inquiry into liability regimes. Hence, our legal analysis will be high-level and comparative to spur debate in diverse jurisdictions. We do not examine the case law of multiple legal systems, as it predates the advent of ML-based AI and is thus not directly relevant for the foreseeability problem we aim to address here [2]. We believe that—despite the appeal of such a quick-fix solution—the *unaltered* application of existing liability rules to AI or a *protectionistically motivated* recourse to strict liability to establish responsibility at any cost are not the correct answers. For this reason, we propose the creation of a system of AI guarantee schemes (AIGSs)—a clear and transparent framework for speedy compensation where a liability suit has uncertain or no prospect of success owing to the unforeseeable nature of the damaging conduct, the (type of) damage itself, or the excessive costs/complexity of the procedure. Mirroring some aspects of financial system guarantee schemes, the AIGSs could function as a second line of defense beyond the ambit, yet complementing the existing system of legal liability. they should be in whole or in part funded by the AI industry.

2 CONDITIONS FOR IMPOSING LEGAL LIABILITY

Legal liability for AI systems could originate from either criminal or civil law. Civil liability can be further divided into contractual, tortious, and statutory liability. We take a comparative legal approach either referring to genuinely transnational sources of law or highlighting common patterns in the law of several jurisdictions.

Contractual liability is premised on a contractual relationship between the parties. We show how it is construed based on the *United Nations Convention on Contracts for the International Sale of Goods (CISG)* [30]—the key international trade law convention governing the international sale of goods, which reflects widely accepted, international commercial best-practices [3]. It adopts a notion of strict liability for breach of contract: Liability for non-performance does not require fault, see Articles 45 and 65 CISG and [3]. However, recognizing that the party in breach cannot control all circumstances leading to non-performance, this unbounded liability is restricted in two ways. First, only *foreseeable* damages can be claimed, (Article 74 CISG). Second, liability is excluded if the *force majeure excuse* (Article 79 CISG) comes into play. Under the force majeure test—where fault becomes relevant—the non-performing party must prove that the breach was caused by an unforeseeable, unavoidable, and insurmountable impediment beyond their control. Thus, in international contract law, the concept of foreseeability determines both the scope of damage claims and the extent to which liability for breach of contract can be established.

Conversely to contractual liability, *tortious liability* can be triggered without a preexisting relationship between the parties. In fact, tort law links wrongdoer and victim—likely total strangers—through the notion of liability to compensate for the harm the former wrongfully inflicted upon the latter [26, 35]. Despite differences in how legal systems construe liability, countries reach similar solutions to similar problems. Our systematic overview is based on [19, 20].

In most jurisdictions, tort law distinguishes between negligence and strict liability, although the extent to which the latter is recognized varies. *Negligence* is a fault-based liability imposed on a tortfeasor that fails to exercise reasonable care, while *strict liability* is negligence's no-fault counterpart, which is typically linked to the existence of a particular source of danger rather than the *conduct* (either action or omission) creating it [15]. A causal relationship between the tortfeasor's conduct and/or the thereby created risks and the victim's harm is universally seen as a minimum condition to shift damage to the tortfeasor and establish their legal obligation for compensation [19]. *Causation* is given if the harm would

not have occurred but for the conduct/risks in question—this is known as the *but for test* or *conditio-sine-qua-non formula* in the legal jargon. However, as discussed below, all legal systems deem such an unrestricted responsibility for all damage that may ensue as a consequence of a conduct unreasonable and employ additional value judgments to confine the scope of liability.

While legal systems often use the terms *wrongfulness, fault, culpability*, and *negligence* interchangeably, they all aim to protect various rights and interests by identifying and preventing potentially harmful and hence wrongful behaviors. Civil law countries have chosen to codify those behaviors in distinct statutory provisions, while in common law systems such standards of conduct have been incrementally developed through case law by defining specific duties of care for different types of torts. Correspondingly, civil law's wrongfulness/fault inquiry focuses on whether the factual elements of a norm have been fulfilled and—if the norm in question establishes fault-based liability—whether the conduct should be qualified as careless under the given circumstances. With common law, responsibility for strict liability torts merely requires proof that a particular harm occurred, it was caused by the defendant's conduct, and the defendant could foresee at least the type of harm that transpired, while in case of negligence torts an additional breach of a particular duty of care by a faulty/negligent conduct is necessary.

Jurisdictions also differ in how they measure fault/negli-gence. The prevalent objective standard of measurement considers a conduct faulty/negligent if it lacks *reasonable* or *ordinary care*, i.e., does not correspond to the way a reasonably prudent person would have acted in the defendant's position. Most strikingly in the US, the negligence standard is an economically charged concept determined by a balancing approach known as the Hand Formula [12]: A conduct is deemed negligent if the *expected* harm—the magnitude of a potential loss (L) adjusted by the probability of its occurrence (P)—outweighs the costs to avoid the harm—the costs of undertaking precautionary measures (B). Put formally, a duty of care is generated where $P \times L > B$ [28, 36]. Other common law jurisdictions rely on this economic logic more covertly and often include additional factors, like the social utility of the conduct, among the balancing criteria, in effect modifying the above formula as follows $P \times L > B + U$, where U stands for social utility.

As pointed out earlier, all jurisdictions reduce the scope of liability delineated solely through causation in two basic ways: They either limit *causation* by using the *theory of adequacy* to exclude liability for objectively unforeseeable damages, or the *scope of liability* by restricting the duty of care to *foreseeable harms*. It follows that, either way, fault based liability can only be imputed for foreseeable harms.

Foreseeability is equally central to strict liability torts, despite the fact that fault plays no role here. Strict liability is imposed on the premise that someone creates a source of danger which is likely to cause harm and, crucially, which said person has the *ability to control*. Yet control implies that both dangerousness and potential harms are recognizable, that is, foreseeable. Similar arguments support the claim that foreseeability is also an essential condition for the imposition of statutory liability: Statutes pre-/proscribe a certain conduct to prevent some risks typically inherent in that behavior from materializing, whereas the scope of a norm cannot reach beyond the limits of foreseeability.

We now turn to a comparative analysis of criminal liability based on [24] and [8]. Criminal punishment presupposes that a particular conduct is criminalized by law, i.e., by statutory provisions. Committing a crime always requires a physical element referred to as *actus reus (guilty act)*. Except for strict liability offenses, where the blameworthiness of a conduct that violates a norm protecting certain societal values is presumed, this must be accompanied by a subjective element referred to as *mens rea* (a.k.a. *culpability*, *fault*, or *blameworthiness*). By contrast to tort liability's objective standards, criminal law measures the defendant's mental state by a predominantly subjective test.

Mens rea is divided into intent (*dolus*), recklessness, and negligence (*culpa*). *Intent* varies in intensity from purposefully committing an offense with the desire to achieve a prohibited result (*direct intent*), acting without such a desire but foreseeing the result as virtually certain (*general intent*), and displaying indifference despite foreseeing a possible harm (*dolus eventualis*). *Recklessness* penalizes behavior that grossly deviates from the standard of conduct of a reasonable person. It is given if an offender is aware of, yet *consciously disregards the substantial and unjustifiable risk* that their conduct will have negative consequences. *Negligence* connotes a behavior that departs from the objective standard of conduct of a prudent person. An *unconsciously* negligent offender *should have been, but wasn't aware of the substantial and unjustifiable risk* of harm, while *conscious* negligence is given if a person *foresees the risk of harm but believes it will not occur*. To justify the imposition of weightier sanctions, criminal law usually requires *gross negligence*, i.e., considerable deviation from the reasonable person standard. It is fulfilled if an individual's actions pose an *obvious risk* to bring about *substantial harm* and the offender has the *ability to take precautionary measures*.

Hence, criminal responsibility likewise presupposes that the offender foresees the potential harms of their conduct. The only exceptions are strict liability and unconscious negligence offenses, but these only entail liability if explicitly criminalized by statutory provisions, which presuppose that the legislator foresees that a conduct may result in harm.

In conclusion, we can observe that foreseeability (and an inherent ability of control) feature prominently among the conditions for imposing any type of legal liability. Admittedly, case law in disparate jurisdictions and legal domains adds a number of convoluted facets to this problem, but for now we would refrain to get into those issues. The important insight at this initial stage is to realize that we face a general legal problem, which spans jurisdictions and legal domains, has potentially severe implications, and consequently needs to be addressed as soon and as widely as possible. On this note, let us now investigate if and to what extent AI is foreseeable/controllable in the sense required by law.

3 FORESEEABILITY: THE MISSING PIECE OF THE AI LIABILITY PUZZLE

Both policymakers and society at large need to be conscious that AI does not *know*, *think*, *foresee*, *care*, or *behave* in the anthropomorphic sense, but applies what could be best described as *machine logic*. ML-based systems—which raise the biggest technical and legal challenges due to their unpredictability stemming from their

independent learning property—do not *know* why a given input should be associated with a specific label (e.g., that a small, red, circular object is a ball), only that certain inputs are *correlated* with that label [23]. They identify outputs based on a set of predefined parameters and probability thresholds through a process fundamentally different from human thinking.

Conventional ML-based systems usually use *human engineered* feature extractors to process raw data in order to receive a suitable representation the system can work with. By contrast, deep learning (DL)-based systems—a neural network-based subgroup of ML approaches—are capable of processing raw data on their own, automatically identifying the right representation they need for classification. They do not just use this one representation, but possess a nested hierarchy of representations obtained by transforming a lower level representation (starting with the raw data) to a higher, more abstract level of representation [21]. This type of machine reasoning always implies a certain probability of failure, where failures tend to occur in—from a human perspective—unexpected ways and may have different reasons. Let us give two examples.

In the first example, the failure is caused by a *bad classifier* as illustrated by [29] in their Husky vs. Wolf experiment. A system trained with 10 wolf and 10 husky pictures was given the task to distinguish between wolves and huskies. On purpose, all wolf pictures had snow in the background but none of the husky pictures. Since snow was a common element in the wolf pictures but was not present in the husky pictures, the system regarded snow as a classifier for wolves. Thus, in the experiment the system predicted huskies in pictures with snow as wolves and vice-versa. The second example shows an *adversarial setting* (essentially an automated attack on a search algorithm to minimize its utility) where the adversary's goal was to create inputs that a DL-based system misclassifies, however, humans do not [33]. Their adversary system manipulated input data by adding what is called *noise* not detectable to human eyes to the original pictures, fooling a DL-based system into classifying a school bus and a pyramid as an ostrich.

Hence, it is conceivable that AI systems and the way they generate failures are too complex to be foreseeable.

4 AI GUARANTEE SCHEMES AS WORK-AROUND

The legal system will need time to incrementally adapt and solve these foreseeability loopholes and other challenges posed by AI. We are probably looking to several decades of deliberation, trial-and-error type of progress in the legal treatment of AI, and inconsistent jurisprudence [27]. So, should we stall the adoption of AI until we can *guarantee* its safety? Or is there a compromise that encourages both reasonable safety and innovation? Again, this problem is not specific to AI but common to all new technologies, and fears about the legal system's ability to rise to certain challenges have prompted a search for alternative solutions in the past.

Examples include *no-fault insurance-based solutions*, which substitute for and eliminate access to the judicial system. Such accident insurance schemes are in place in diverse countries and fields like occupational, medical, and all types of personal injuries. Dispensing with the need to examine how the damage occurred, they guarantee victims fast compensation at lower administrative costs compared to litigation. But, they promote carelessness and, due to financial constraints, typically only offer partial compensation through the introduction of arbitrary restrictions unlike the judicial system, which fully compensates victims after successful litigation. Measures to alleviate these weaknesses—such as making the amount of compensation conditional on the specific circumstances under which the damage occurred or granting insurers rights of recourse against tortfeasors—help to provide a fairer and more equitable compensation, but do so at the cost of speed and increased costs due to the necessary legal inquiry into causation.

Another approach, analyzed by [27], is setting up *victim compensation funds*—e.g., the 9/11 Victim Compensation Fund and the Gulf Coast Claims Facility established after the Deepwater Horizon disaster—which exist parallel to rather than in lieu of the judicial system. Such funds' objectives include relieving pressure on courts, supporting ailing industries, or simplifying and expediting compensation processes. They may be established either as quasi-judicial or non-judicial funds. *Quasi-judicial funds* are administered by the judicial system or a public agency and financed by taxes or fines imposed on a selected group of individuals or organizations, who would assume a defensive position in litigation. *Non-judicial funds* are divided into three sub-categories: *Public funds*, administered and at least partially funded by government or an entity with government authority, *private funds*, administered and funded by private organizations, and *charitable funds*. These are also privately administered and funded by private donations, yet are distinct from the other three types of funds in two respects. First, their only purpose is to minimize administrative and logistical burdens of distributing donations rather than providing an alternative to litigation. Second, they provide flat compensation awards without recourse to tort law to determine eligibility for compensation. One advantage of victim compensation funds over conventional litigation is flexibility, since their status, funding, administration, and processes are designed with a particular set of circumstances in mind. They also tend to be faster and more efficient and cost-effective than the tort system. However, establishing funds potentially involves much higher administrative burdens compared to the judicial system. Unlike conventional litigation, they typically do not provide transparency and publicity, although this may be important to victims.

Pearl recommends the creation of a fund for AV crash victims in the US until the legal system catches up with AI innovation. She proposes a quasi-judicial fund administered by the National Highway Transportation Safety Administration (NHTSA) and funded by taxes on the sale of AVs to be paid by sellers and purchasers. Both buyers and sellers should contribute, as both groups would benefit from the introduction of AVs. She envisages a voluntary participation both for victims (requiring them to file a claim with the fund and waive their right to litigation upon acceptance of the compensation award) and AV manufacturers (under the condition of paying their share of the AV sales tax and participating in data-sharing and design improvement programs). The fund should only cover human injuries and fatalities. Compensation should be full and automobile insurance companies (whose subrogation rights would be extinguished where victims accept compensation awards) should be allowed to seek reimbursement from victims' compensation awards to recover prior insurance payouts.

Our proposal to create a system of *AIGSs* is inspired by the various types of guarantee schemes (most notably deposit guarantee, insurance guarantee, and investor compensation schemes) used in the financial system (hereinafter FGSs.) FGSs are usually sectorally configured, at least partially industry-funded, and sovereign-backed guarantee funds. Together with other arrangements—like lender or market maker of last resort support from central banks—they make up the heterogeneous group of *financial system guarantees*. Broadly speaking, these guarantees (a.k.a. *financial system safety net*) are designed to provide assurance to those involved in financial transactions that their claims against their counterparties will be met even in the event of a major liquidity shock or failure of the latter. Heavily expanded in the wake and after the global financial crisis, they are a widely used and successful model to safeguard financial stability by preserving confidence in the financial system in times of stress [5, 31].

This powerful feedback-loop between the extent of uncertainty and the level of trust is a central determinant in shaping any market, AI being no exception. So, to foster confidence, the AIGSs should provide a transparent, predictable, and reliable alternative funding mechanism outside of the scope of the legal liability system to compensate aggrieved parties. Compensation should be available in a contractual, tort, or, as appropriate, criminal context in cases where legal liability cannot be established due to the lack of foreseeabilty of an AI performance failure and/or the resulting harm. AIGSs could also shore up the legal system in the face of the anticipated uncertainty and complexity of AI-related litigation. Because such difficulties are likely to occur worldwide and in all domains impacted by AI, our proposal is geared to the global context, taking a country- and domain-neutral approach. Our goal is to spark a high-level, conceptual debate to inform future policy initiatives.

Governance arrangements of any guarantee scheme are strongly dependent on the broader governance structures adopted in industries to which they are linked. Given the preliminary stage of discussion on AI governance in virtually any domain and country, it is relatively hard to define robust design criteria. Nevertheless, based on a survey of international practices with respect to FGSs [5] and Pearl's above recommendation we will sketch out an initial set of principles to guide future deliberations on this issue.

Nature of the scheme: Beyond providing predictability regarding compensation, we see AIGSs as integral parts of the broader domestic and eventually global AI governance frameworks, which pursue the overarching objective of ensuring that the development and adoption of AI is beneficial to humanity. One facet of that endeavor is to incentivize AI innovators to employ responsible and safe practices, but the funds could also further other policy objectives, such as mitigating AI's inequality-aggravating impacts by redistributing some of the costs and benefits of AI innovation. In light of these strong public policy implications, quasi-judicial funds do indeed seem best suited to function as AIGSs.

Administration: There is growing consensus about the necessity of a global AI governance framework [6, 17, 25]. Presently, however, AI innovation and implementation is outpacing policymakers' regulatory and oversight capabilities, and countries' focus is restricted to tackling the most pressing issues across diverse policy domains, without much regard to cross-sectoral consistency. Policy domains have their established regulatory frameworks, traditions, and specific difficulties, requiring specialized expertise and sector-specific regulation [37]. This and nascent national practices suggest that AI governance will initially be structured in a domain-specific fashion with existing agencies taking on AI-related regulatory functions. Given the need for speedy policy response, this is a commendable approach at least on the interim, until more research can be done on the optimality of governance arrangements. For FSGs, [5] identifies six key governance objectives stressing that governance arrangements should (1) establish clear lines of responsibility avoiding duplication of regulatory mandates, (2) eliminate avenues for conflicts of interests, (3) minimize the administrative costs of the fund and (4) compliance burdens for industry, (5) where appropriate, involve industry stakeholders harnessing their expertise, and (6) provide adequate incentive structure for regulatory authorities.

These observations furnish strong arguments to house AIGSs within domain-specific agencies at least until we can explore alternatives. In the meantime, we strongly encourage the international community to keep up efforts towards setting up a global AI governance framework—preferably involving some element of self-regulation to benefit from multifaceted expertise and ensure a truly dynamic whole-of-society dialogue. Once up and running, such cross-jurisdictional governance arrangements could justify a transnationally organized AIGS system.

Coverage: As noted by [5], fund coverage design inevitably involves wrestling with tradeoffs between the conflicting objectives of efficiency, equity, and minimum complexity/cost. The costs of guarantee schemes are not restricted to the amount of compensation paid out, but also include potentially significant administrative and compliance costs (e.g., costs of establishment and ongoing operation of schemes, dispute resolution mechanisms) and much less obvious indirect costs to society in the form of moral hazard and related behavioral problems. The appropriate balance between different objectives is typically sector-dependent and tools like coverage limits, coinsurance, and means testing are among those employed to find a suitable configuration. In view of guarantee schemes' role as safety net, i.e., a sort of back-up solution, they should ideally only step in to compensate substantial losses. Given the abundance of unknown variables in this respect, we would refrain from offering any specific recommendation at this time.

Participation: In theory, participation in guarantee schemes may be either voluntary or compulsory. FGSs typically foresee compulsory participation to avoid problems of adverse selection, i.e., disproportionate representation of the least reliable institutions in funds. This argument also holds for AI innovators' and manufacturers' recourse to AIGSs, suggesting that compulsory participation may be preferable. This could additionally be justified by AIGSs' intended rational as a tool to regulate the AI industry's incentive structure, while potentially also pursuing other policy objectives.

Funding and pricing: Guarantee schemes involve a redistribution of losses: Certain stakeholders foot the bill to alleviate pressure on others. Striking a level of redistribution that stakeholders perceive as fair is key to ensure guarantee funds' acceptance and efficiency. Funding relates to the timing and rate of contributions, as well as the base of funding. Pricing determines contributors' relative share. With FGSs, funding and pricing considerations should pursue four goals [5]: (1) cost efficiency (minimize administrative costs), (2)

competitive neutrality (equitable treatment of contributors of similar characteristics), (3) stability (predictable and broadest possible funding base), and (4) allocative efficiency (eliminate moral hazard incentives).

In terms of the timing of funding, funds can be either pre-funded (contributions are paid into and managed by the fund) or post-funded (contributors incur contingent liabilities and are only required to pay into the fund after the guarantee triggering event), or a combination of both. Pre-funding usually implies greater stability and credibility that funds are readily available in crisis. It is also conducive to a higher acceptance of risk-sensitive pricing, typically perceived as fair, and requires less financial back-up by the public purse. On the down side, pre-funding may lead to higher than warranted contributions due to the uncertainty of triggering events' occurrence. It may also create moral hazard incentives, raise issues around controlling the size of the contribution pool, and be less cost efficient than post-funding. Post funding, on the other hand has a pro-cyclical impact, in that it imposes a burden on contributors after a guarantee event, compounding their financial difficulties.

Funding base-related questions revolve around the relative ratio of public and private funding, whether to establish several domain-specific schemes or one cross-sectoral fund, and the basis for calculating contributions. The pros of domain-specific schemes include cost efficiency, competitive neutrality, sensitivity to domain specific characteristics, and avoidance of cross-subsidies. But they are less financially stable, have a restricted ability to realize diversification benefits, and may face transition problems due to structural changes in the organization of contributing entities.

Pricing choices aim to strike an acceptable balance between simplicity and efficiency. The latter is promoted by differential, risk-sensitive contributions, which are better at combating moral hazard and ensuring equitable treatment of contributors, but complex to implement. The alternative is to require uniform, flat-rate contributions, which excel in simplicity, transparency, and involve low implementation costs.

Applying these insights to AIGSs, systemic crises with the potential to deplete FGSs and necessitate state involvement are unlikely in the AI context. Better feasibility of risk-sensitive pricing and the likelihood that industry would perceive it as the fairer option are still strong arguments in favor of pre-funding, whereby post-funding—even in an auxiliary form—will probably not be necessary. Unless specific policy considerations dictate otherwise, the funding base should be restricted to private contributions from the AI industry—the group whose incentive structure it aims to target—without involving public funds or contributions from AI users. Recalling our above recommendation for domain-specific AI governance arrangements, funding should be organized on a sectoral basis. Contributions should be calculated taking due account of domain-specific criteria based on, e.g., the estimated amount of compensation awards obtainable in litigation. Given the importance of the perceived fairness of schemes' redistributive effects, we strongly favor risk-sensitive pricing arrangements. We also call for considering risk management techniques from the financial regulatory domain to gauge contributors' risk to FGSs as a possible model to overcome hurdles of complexity.

Compensation process: In terms of the process by which victims and otherwise aggrieved parties may obtain compensation, Pearl's simple, non-adversarial approach—requiring claimants to file a claim with an AIGS outlining the grounds for a compensation award and waiving their right to litigate upon acceptance of the award—coupled with appropriate appeal and dispute resolution mechanisms would presumably be suitable for most AI domains.

5 CONCLUSION

With an eye on the primary objective pursued by AI innovation—enhancing inclusive economic and social welfare across the globe—this paper has exposed weaknesses in the existing system of legal liability and recommended the creation of a system of AIGSs. As a predictable and transparent framework for swift compensation outside of the purview of legal liability, the AIGSs would provide legal certainty in dealing with AI-related liability issues without violating existing liability doctrines, and could also assume a broader role within the overall regulatory framework structuring our economies. They would induce a legal environment that fosters safe and responsible AI innovation and adoption in society, facilitating a smooth transition into an AI-driven society.

REFERENCES

[1] Abrahams, N., Azzopardi, M., Blackwood, V., Erdélyi, G., Erdélyi, O. J., Guihot, M., Lea, G., Liddicoat, J., Matthew, A., Freehills, H. S., Suzor, N., and Commission, A. H. R. Emerging Responses and Regulation. In *The Effective and Ethical Development of Artificial Intelligence: An Opportunity to Improve Our Wellbeing*, T. Walsh, N. Levy, G. Bell, A. Elliott, J. Maclaurin, I. M. Y. Mareels, and F. M. Wood, Eds. Australian Council of Learned Academies, 2019, pp. 132–153. Report for the Australian Council of Learned Academies, www.acola.org.

[2] Barfield, W. Liability for Autonomous and Artificially Intelligent Robots. *Paladyn, Journal of Behavioral Robotics 9*, 1 (2018), 193–203.

[3] Brunner, C. *Force Majeure and Hardship under General Contract Principles*. Kluver Law International, 2009.

[4] Calo, R. Law and Technology: Is the Law Ready for Driverless Cars? *Communications of the ACM 61*, 5 (2018), 34–36.

[5] Davis, K. Study of Financial System Guarantees, 2004.

[6] Erdélyi, O. J., and Goldsmith, J. Regulating Artificial Intelligence: Proposal for a Global Solution. In *AIES'18* (2018), pp. 95–101.

[7] European Commission. Ethics Guidelines for Trustworthy AI, 2019. 8 April 2019.

[8] Fletcher, G. P. *Rethinking Criminal Law*. Oxford University Press, 2000.

[9] G20. G20 Ministerial Statement on Trade and Digital Economy, 2019. Meeting of 8 and 9 June 2019.

[10] Gerstner, M. E. Liability Issues with Artificial Intelligence Software. *Santa Clara Law Review 33*, 1 (1993), 46–51.

[11] Greenblatt, N. A. Self-driving Cars and the Law. *IEEE Spectrum 53*, 2 (2016), 46–51.

[12] Hand, J. L. United States v. Carroll Towing Co., 159 F.2d 169 (2d Cir. 1947), 1947.

[13] Hubbard, F. P. Allocating the risk of physical injury from "sophisticated robots": Efficiency, fairness, and innovation. In *Robot Law*, R. Calo, A. M. Froomkin, and I. Kerr, Eds. Edward Elgar Publishing, 2016, pp. 25–50.

[14] Karnow, C. E. A. The Encrypted Self: Fleshing Out the Rights of Electronic Personalities. *The John Marshall Journal of Information Technology and Privacy Law 13*, 1 (1994), 1–16.

[15] Karnow, C. E. A. The application of traditional tort theory to embodied machine intelligence. In *Robot Law*, R. Calo, A. M. Froomkin, and I. Kerr, Eds. Edward Elgar Publishing, 2016, pp. 51–77.

[16] Kingston, J. K. C. Artificial intelligence and legal liability. In *Research and Development in Intelligent Systems XXXIII*, M. Bramer and M. Petridis, Eds. Springer International Publishing, 2016, pp. 269–279.

[17] Koene, A., Richardson, R., Hatada, Y., Webb, H., Petel, M., Reisman, D., Machado, C., Violette, J. L., and Clifton, C. A governance framework for algorithmic accountability and transparency, 2018. EPRS/2018/STOA/SER/18/002.

[18] Korinek, A., and Stiglitz, J. E. Artificial Intelligence and Its Implications for Income Distribution and Unemployment. In *The Economics of Artificial Intelligence: An Agenda*, A. K. Agrawal, J. Gans, and A. Goldfarb, Eds. University of Chicago Press, 2019.

[19] Koziol, H. *Basic Questions of Tort Law from a Germanic Perspective*. Atlasbooks Dist Serv, 2012.

[20] Koziol, H., and Askeland, B. *Basic Questions of Tort Law from a Comparative Perspective*. Jan Sramek Verlag, 2015.

[21] LeCun, Y., Bengio, Y., and Hinton, G. E. Deep learning. *Nature 521*, 7553 (2015), 436–444.
[22] Liechtung, J. The Race is On! Regulating Self-Driving Vehicles Before They Hit The Streets. *Brooklyn Journal of Corporate, Financial & Commercial Law 12*, 2 (2018), 389–413.
[23] Lipton, Z. C. The Mythos of Model Interpretability. *Queue 16*, 3 (June 2018), 30:31–30:57.
[24] Marchuk, I. *The Fundamental Concept of Crime in International Criminal Law: A Comparative Law Analysis.* Springer Berlin Heidelberg, 2014.
[25] OECD. Recommendation of the Council on Artificial Intelligence, 2019. C/MIN(2019)3/FINAL.
[26] Oliphant, K. Basic Questions of Tort Law from the Perspective of England and the Commonwealth. In *Basic Questions of Tort Law from a Comparative Perspective*, H. Koziol and B. Askeland, Eds. Jan Sramek Verlag, 2015, pp. 355–430.
[27] Pearl, T. Compensation at the Crossroads: Autonomous Vehicles and Alternative Victim Compensation Schemes. In *ITS European Conference* (2018).
[28] Posner, R. A. A Theory of Negligence. *The Journal of Legal Studies 1*, 1 (1972), 29–96.
[29] Ribeiro, M. T., Singh, S., and Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *KDD'16* (2016), ACM, pp. 1135–1144.
[30] Rome statute of the international criminal court, 1998.
[31] Schich, S., and Kim, B.-H. Guarantee Arrangements for Financial Promises: How Widely Should the Safety Net be Cast? *OECD Journal: Financial Market Trends 2011*, 1 (2011), 201–235.
[32] Stiglitz, J. E. Rewriting the Rules of the American Economy: An Agenda for Growth and Shared Prosperity, 2015. Roosevelt Institute.
[33] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. Tech. rep., CoRR, 2013. abs/1312.6199.
[34] Vladeck, D. C. Machines Without Principals: Liability Rules and Artificial Intelligence. *Washington Law Review 89*, 117 (2014), 117–150.
[35] Weinrib, E. J. *The Idea of Private Law.* Oxford University Press, 2012.
[36] White, B. A. Risk-Utility Analysis and the Learned Hand Formula: A Hand That Helps or a Hand That Hides? *Arizona Law Review 32*, 1 (1990), 77–136.
[37] Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., and Schwartz, O. AI Now Report, 2018.