Journal of Artificial Intelligence Research 70 (2021) 1309-1334 Submitted 12/2020; published 04/2021

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

Olivia J. Erdélyi OLIVIA.ERDELYI@CANTERBURY.AC.NZ
University of Canterbury, School of Law
Christchurch 8140, New Zealand
Soul Machines
Auckland 1010, New Zealand

Gábor Erdélyi GABOR.ERDELYI@CANTERBURY.AC.NZ
University of Canterbury, School of Mathematics and Statistics
Christchurch 8140, New Zealand

Abstract

Confidence in the regulatory environment is crucial to enable responsible AI innovation and foster the social acceptance of these powerful new technologies. One notable source of uncertainty is, however, that the existing legal liability system is unable 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 difficult to calculate with confidence, 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 frameworks, 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 it to be at least partially industry-funded. Funding arrangements should depend on whether it would pursue other potential policy goals aimed more broadly at controlling the trajectory of AI innovation to increase economic and social welfare worldwide. Because of the global relevance of the issue, rather than focusing on any particular legal system, we trace relevant developments across multiple jurisdictions and engage in a high-level, comparative conceptual debate around the suitability of the foreseeability concept to limit legal liability. The paper also refrains from confronting the intricacies of the case law of specific jurisdictions for now and—recognizing the importance of this task—leaves this to further research in support of the legal system's incremental adaptation to the novel challenges of present and future AI technologies.

1. Introduction

With proliferating AI-human interactions, issues around the civil and criminal liability for AI systems have moved to the forefront of legal 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 is used 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 touches upon 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? Well aware that courts and policymakers will soon have to come up with satisfactory answers, a growing number of papers has taken a first crack at examining the topic from various perspectives. The result is a landscape of conflicting accounts on how best to go about AI liability and the legal system's overall ability to adapt to this latest wave of technological innovation.

Some further the debate by synthesizing the relevant literature on selected aspects of civil and criminal liability (Kingston, 2016; Perc, Ozer, & Hojnik, 2019). Among those having faith in the existing system's adequacy to deal with AI liability issues is Hubbard (2016), who—tacitly invoking the famous Hand formula (Judge Learned Hand, 1947)—concludes that the current US system of contractual and tort liability strikes a fair and efficient balance between ensuring safety and incentivizing innovation. Consequently, they see no reason to apply different metrics to compensating physical injury inflicted by *sophisticated robots* (which they define as having some degree of connectivity, autonomy, and potentially machine-learning (ML) ability).

Given the apparent imminence of the topic, quite a few papers revolve around AVs. Liechtung (2018) urges for timely adjustment of regulation and oversight mechanisms to prepare for the impending mass-release of AVs. They also stress the importance of clarity and predictability of the legal liability regime—whatever liability rules are chosen—so those involved in the development, production, and distribution of AVs can better assess their risk exposure.

Several commentators argue for subjecting AVs or AI more broadly to strict liability—commonly some type of products liability regime (Gerstner, 1993; Liechtung, 2018). An interesting recent idea in this realm has been put forward by Vladeck (2014), advocating a strict liability regime entirely detached from notions of fault—in essence a court-implemented insurance system. In a practical, goal-oriented, if slightly doctrinally inconsistent approach, they propose to simply infer liability from negative outcomes to overcome situations where it is impossible to establish fault. Doing so, they hope to create a more cost-efficient, equitable, and predictable liability regime, which provides a safe and stimulating environment for innovation, better protection to blameless parties, and fairer cost-spreading among affected parties. As a way of achieving the latter, they contemplate abandoning the current practice of treating AV liability as an agency question (Köhler, 2018) and conferring legal personality on AVs coupled with a compulsory self-insurance instead. Relatedly, Karnow (1994) advocates an *electronic personality* for autonomous robots (by which they mean those with an ML component) to enable the legal system to hold them directly liable under tort law. Whether making AI systems legal persons—thereby eliminating any direct human responsibility and oversight of their operation and impact on the environment—is a sound idea is a controversial issue in itself, with numerous ethical, design, and legal considerations speaking against it (Institute of Electrical and Electronics Engineers, 2019; Bryson, Diamantis, & Grant, 2017).

A contrasting view reveals concerns about potentially 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 to AV incidents (Greenblatt, 2016). They maintain that equal treatment of AVs and those under human guidance in this manner would also result in a higher degree of legal certainty spurring innovation—a common theme supporting all the above views—and allow for the operation of market-based incentives such as reputational concerns.

Finally, Karnow (2016) expresses doubts as to whether any of the classic United States 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 of these doctrines, however, due to complex non-linear interactions between intricate robots and their equally convoluted, unpredictable environment, neither robots' actions nor the potential harms they may cause are necessarily foreseeable in the sense required by law. Regarding AVs and autonomous robots and mostly in the context of United States tort law, other commentators have voiced similar concerns about potential liability gaps and the implications of a resulting overall uncertainty surrounding the legal liability of AI systems. They confirm Karnow's observation about the centrality of foreseeability in limiting legal liability but concede that emergent behavior—i.e., behavior contingent on the interaction of a system's elements rather than the elements themselves—exhibited by some systems may trigger genuinely unforeseeable categories of harm (Barfield, 2018; Calo, 2018). 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. As explored in economics and law and technology literature, the presence of uncertainty—especially coupled with the absence of individuals' ability to insure themselves against it—can significantly inhibit innovation and the adoption of new technologies, in extreme cases reaching as far as shutting down entire emerging markets (Arrow, 1962; Pearl, 2018).

In this paper, we restrict the focus of the above sketched AI liability debate by analyzing only the foreseeability concept's ability to serve as a means to limit and attribute legal liability. At the same time, however, we will also move this important discussion beyond United States tort law and embodied AI systems or particular AI applications—indeed beyond any national analysis and law in general, for the following reasons:

AI is just one of the most recent waves of technological innovation (sometimes referred to as the fourth industrial revolution) all of which have fundamentally impacted our societies and economies. Due to its rapid pace of development, massively transformative nature, and other changes—most notably globalization—our world has undergone, AI is anticipated to affect humanity and our environment even more intensely. Recognizing this, there are major national, regional and international AI strategies and policy initiatives underway, which aim to forge an innovation-friendly, enabling regulatory environment, capturing benefits and minimizing potential risks AI may bring (G20, 2019; Organisation for Economic Co-operation and Development (OECD), 2019; European Commission High-Level Expert Group on Artificial Intelligence, 2019; European Commission, 2020c; Abrahams et al., 2019). All these initiatives converge on the point that successful societal adoption of AI—like any other form of technological innovation—requires trust on the part of society. Trust, in turn, 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 of the technology must be assured that its use is reasonably safe. Without such trust and certainty, innovation in emerging technologies is likely to be severely stifled, hampering economic growth and welfare (Arrow, 1962; Pearl, 2018). Certainty itself flows from a safe, transparent, and flexible regulatory environment that supports innovation. The 2018 fatal Uber AV crash in Tempe, Arizona, and its aftermath, on the other hand, provides a warning example of how regulatory failures coupled with human errors can shatter trust and threaten markets in an emerging technology (National Transportation Safety Board (NTSB), 2018; Templeton, 2019). Yet, as explained below, designing any aspect of the nascent AI regulatory framework—such as an adequate legal liability regime—is neither a purely legal nor an exclusively national enterprise.

From an economic perspective, the regulatory frameworks that structure our economies together with market imperfections crucially determine the extent to which society benefits from technological innovation (Stiglitz, 2015). It is established wisdom that in our reality of imperfect markets, technological innovation is not necessarily Pareto-improving. On the contrary, in the absence of cleverly devised and potentially substantial redistributive measures, it can actually aggravate inequality and decrease overall welfare (Korinek & Stiglitz, 2019). Stiglitz (2015) also shows that 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 admittedly part of. This argument advocating a holistic regulatory stance also holds true more generally when it comes to optimally aligning different regulatory objectives within a broader regulatory system.

The above cited key international AI policy documents, and—based on a review of relevant international relations literature— Erdélyi and Goldsmith (2018, 2020) 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, can be far more effectively regulated in an internationally coordinated manner—be it in the form of truly transnational policy initiatives or national measures that display at least some degree of coordination. The reason is that without coordination, domestic approaches will inevitably be fragmented. This not only invokes inefficiencies and tensions in international policymaking, but also negatively affects domestic regimes, shattering both national and international actors' faith in the viability of national regulatory approaches.

These arguments combined call for a holistic, multidisciplinary, and transnational perspective, forbidding an isolated legal or nationally focused analysis of liability regimes. Hence, Section 2 will start with a high-level, comparative legal analysis, which should help the reader to navigate the differences in terminology and/or legal approaches that persist between the two main legal traditions—common law and civil law systems—and even across countries belonging to the same tradition. It will explain why the notion of foreseeability is central to attributing legal liability in all legal systems and highlight a potential conceptual problem related to its suitability to constrain the legal liability of AI systems. We argue that foreseeability is common to all types of legal liability irrespective of the area of law they originate from, and may raise attribution problems in relation to the actions of any embodied or disembodied AI system, provided it uses certain types of ML models. Note that although under the current state of technology these problems arise in connection with certain (not all) ML-based systems, in the future they are equally conceivable in relation to other types of AI systems. Therefore—as outlined in Section 3 in more detail—we understand the term *AI system* as defined by the Organisation of Economic Co-operation and Development (OECD) in the OECD Principles on Artificial Intelligence (Organisation for Economic Co-operation and Development (OECD), 2019), as this allows for the inclusion of any present and future AI technologies that may pose similar challenges. We use the more generic term *AI* in effect interchangeably, referring to a diverse set of technologies that it encompasses at any given time. Our aim is to spur further legal debate in diverse jurisdictions. Examining the case law of multiple legal systems is, however, beyond the scope of the paper, not least because—as Barfield (2018) points out with respect to US case law dealing with robots—cases that could serve as precedents date back to prior to the advent of ML-based instantiations of AI and consider non-autonomous systems. Consequently, they are not directly relevant for the foreseeability problem we aim to address here.

Section 3 will recommend an approach to tackle current difficulties policymakers face in trying to define AI, and also engage in a short technical analysis of the current spectrum of ML-based AI systems to illustrate that a subset of them does not fulfill the foreseeability requirement.

As noted above, both prolonged legal uncertainty and unduly restrictive policies stall innovation. Hence, Section 4 will first suggest a minor amendment to the existing legal liability regime—as emerging based on our comparative legal analysis. This reflects our belief 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 with a view to establish responsibility at any cost are not the correct answers for three reasons: First, ignoring that those rules have been tailored to different circumstances and may hence be inappropriate for AI, they contravene the delicately balanced objectives of the legal liability system. Second, they inhibit AI innovation by adopting an unduly punitive approach. Third, undue resort to strict liability merely circumvents foreseeability and fault problems in a dogmatically inconsistent manner rather than remedying them.

Section 5 will then discuss our thoughts on the simultaneous creation of a system of AI guarantee schemes (AIGSs). This would essentially be an insurance mechanism against uncertainty: A clear and transparent framework for speedy compensation in cases 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 and/or complexity of the procedure. Mirroring some aspects of financial system guarantee schemes—which form an integral part of the financial safety net and crucially contribute to maintaining trust in the financial system—the AIGSs could function as a second line of defense beyond the ambit, yet complementing the existing system of legal liability. Depending on the AIGSs' designated role within the broader regulatory frameworks structuring our economies and the policy goals they pursue, they should be in whole or in part funded by the AI industry.

Summary of Contributions:

- Comparative legal analysis of the foreseeability concept's role and adequacy in constraining and attributing liability in the context of AI systems across *all* legal domains.

- Multidisciplinary arguments to show the economic and political costs of failure to solve the identified foreseeability problem in a timely and internationally coordinated manner.

- Discussion and concrete examples to show that *there exist* AI systems that do not fulfill the legal foreseeability requirement.

- Recommendation on how to approach definitional difficulties in AI policy and regulatory debates.

- Proposal to amend existing legal liability regimes to account for AI's social utility.

- Proposal to set up an insurance mechanism against uncertainty in the form of a system of AIGSs—based on financial regulatory best practices—to minimize legal uncertainty and foster safe and responsible AI innovation.

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 according to the particular field of law from which the liability emanates. Conscious of the importance of striving for a globally consistent treatment of AI-related issues from the outset (Erdélyi & Goldsmith, 2018, 2020), we would like to once more note our intention to take a comparative legal approach either referring to genuinely transnational sources of law or highlighting common patterns in the law of several jurisdictions. In so doing, we hope to provide an analysis that resonates with the international community and may inform both domestic and international policy initiatives in this space.

Contractual liability is premised on a contractual relationship between the parties. We illustrate how it is construed based on the *United Nations Convention on Contracts for the International Sale of Goods (CISG)* (United Nations (UN), 1980)—the key international trade law convention governing the international sale of goods. The CISG is viewed as prima facie evidence for general contract principles reflecting widely accepted, international commercial best-practices making up the core of a broad set of legal systems (Brunner, 2009). As such, it is not only derived from and influencing national legal systems, but also crucially guides international commercial arbitral tribunals, which will likely be heavily involved in the adjudication of AI-related commercial disputes in the coming years.

In common law tradition, the CISG adopts a notion of strict liability for breach of contract: The party failing to perform its contractual obligations is liable for non-performance regardless of whether fault can be established on their part, see Articles 45 and 65 CISG, and Brunner (2009). 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, i.e., loss the non-performing party has or "ought to have" foreseen as a possible consequence of the breach based on information they did or "ought to have" known when concluding the contract (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. A foreseeable impediment is defined as one that parties could "reasonably be expected to have taken [...] into account." The foreseeability requirement determines the contractual risk allocation through an implicit, but rebuttable assumption that the party in breach assumes risk for the occurrence of foreseeable circumstances. For a good overview, see the book by Brunner (2009). 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 irrespective of whether the parties have a preexisting relationship. In fact, in line with the *principle of corrective justice*, 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 (Oliphant, 2015; Weinrib, 2012). A comparison of different legal systems yields a somewhat confusing picture as to how liability is construed, but at the end of the day, countries reach similar solutions to similar problems. Our attempt to outline a systematic overview is based on the works of Koziol (2012) and Koziol and Askeland (2015).

Starting with the common traits, in most jurisdictions, tort law distinguishes between negligence and strict liability, although the extent to which the latter is recognized varies considerably. *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 (Karnow, 2016). Pursuant to the *principle of bilateral justification* characterizing private law, 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 (Koziol, 2012). *Causation* is given if the harm would not have occurred but for the conduct or 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 some conduct unreasonable and employ additional value judgments to confine the scope of liability. This is where the similarities stop and the inconsistencies start.

First, there is a disturbing amount of conceptual inconsistency across various legal systems with regard to the—often interchangeably used—terms wrongfulness, fault, culpability, and negligence. As far as wrongfulness and fault are distinguished, *wrongfulness*—a term widely used in Germanic countries—is an objective concept that refers to misconduct, i.e., a conduct that is somehow incorrect in the eyes of the law, whereas *fault* is a subjective notion which serves to assign blame to a certain misconduct. Systems where such a distinction does not exist follow essentially the same logic by uniting objective and subjective considerations under the umbrella of a single term—fault in France—or resorting to additional helping mechanisms—e.g., duties of care in English and United States law. English law views fault as an element of wrongfulness, yet tends to use both concepts interchangeably. United States law makes no distinction at all and prefers the term *culpability* or fault over wrongfulness. *Negligence*—popular especially in common law jurisdictions—either means the fault-based class of tort liability by contrast to strict liability or is used as a synonym for fault and culpability.

To make sense of this, it is helpful to acknowledge that all legal systems aim to protect various rights and interests by identifying and preventing potentially harmful and hence wrongful behaviors. There is, however, a key difference in how civil and common law jurisdictions go about this: Following a more systematic approach, civil law countries have chosen to codify such behaviors—which form the basis of wrongfulness—in distinct statutory provisions. In common law systems, on the other hand, such standards of conduct have been incrementally developed through case law by defining specific duties of care for different types of torts. Correspondingly, to hold a defendant liable for damages caused by their conduct, civil law's wrongfulness or fault inquiry focuses on whether the factual elements of a norm have been fulfillled and—if the norm in question establishes fault-based liability—whether the conduct should be qualified as careless under the given circumstances. Common law approaches torts in a slightly different, yet in terms of the outcome essentially similar manner: Responsibility for strict liability torts merely requires proof that a particular harm occurred, that said harm was caused by the defendant's conduct, and that the defendant could foresee at least the type of harm that transpired. In case of negligence torts, an additional breach of a particular duty of care by a faulty or negligent conduct is necessary.

Jurisdictions also differ in how they measure fault and negligence. The prevalent objective standard of measurement considers a conduct faulty or 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 United States, the negligence standard is an economically charged concept determined by a balancing approach—essentially an economic cost-benefit analysis—known as the already mentioned Hand formula (Judge Learned Hand, 1947): 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 burden of undertaking precautionary measures (B). Put formally, a duty of care is generated where $P \times L > B$ (Posner, 1972; White, 1990). 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 such that $P \times L > B + U$, where U stands for social utility. This objective approach is usually justified with reference to the exorbitantly high administrative costs of determining each defendant's abilities on an individual basis, the observation that the tortfeasor's abilities have no bearing on how their actions affect others, the endeavor to reinforce people's moral responsibility, or that the law needs to define average standards of conduct in pursuit of general welfare. Proponents of a subjective approach criticize that this amounts to an imposition of strict liability in cases where the defendant's abilities are below average. There is little practical difference between the two approaches, as ultimately both require courts' discretionary judgment on whether it is reasonable to impose liability on a case-by-case basis.

As pointed out earlier, all jurisdictions reduce the scope of liability delineated solely through causation. Here again, approaches and terminology are confusingly inconsistent both across jurisdictions and different points in time, but restrictions are achieved in two basic ways: By limiting either *causation* or the *scope of liability.* The first technique—limiting causation—works with an unbounded notion of fault, imputing liability for *all* damages caused by a conduct. At the same time, however, it treats causation as a *normative* rather than natural concept: Whether causation is deemed given depends not only on the existence of a cause-effect relationship, but also on additional value judgements. This approach allows reliance on concepts like the *theory of adequacy* to exclude liability for *atypical* or *remote* damage, i.e., damage stemming from an entirely coincidental, objectively unforeseeable interplay of circumstances, which the tortfeasor could not have possibly controlled. The second method—limiting the scope of liability—conceives of causality in the natural sense of the term and—based on prediction theory—limits the scope of liability by restricting the duty of care to *foreseeable harms*, i.e., those the defendant should have actually been capable to avoid (Karnow, 2016). It follows that, either way, fault based liability can only be imputed for foreseeable harms.

Perhaps less intuitively, foreseeability is equally central to strict liability torts, despite the fact that fault plays no role here. To understand why, consider that 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. United States doctrines of strict liability include ultrahazardous activity and three types of products liability. In civil law systems, a number of specific statutory provisions prescribe non-fault based liability for keepers of, e.g., animals and motor vehicles, and other hazardous activities. Jurisdictions with a strict-liability-averse stance, such as England, solve such cases over negligence, but they tend to stretch duty of care requirements so far that liability becomes virtually inevitable—yet another example of similar results achieved by seemingly distinct approaches.

Similar arguments support the claim that foreseeability is also an essential condition for the imposition of statutory liability: Statutes pre- or 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. Instruments like the *protective purpose theory* in some European legal systems or the *harm-within-the-risk rule* in the United States serve the purpose to limit liability for breach of statutory provisions based on this logic.

Turning to criminal liability, we now outline the basic requirements for establishing criminal responsibility based on the comparative analyses of Anglo-American, continental, and international criminal law provided by Fletcher (2000) and Marchuk (2014). Pursuant to the *legality principle*—a central moral principle of criminal law expressed by the Latin term *nullum crimen sine lege (no crime without law)*—criminal punishment typically presupposes that a particular conduct is criminalized by law, i.e., penalized behavior and potential sanctions—the severity of which reaches well beyond those imposed under civil law—are clearly laid down in statutory provisions. Criminal law pursues primarily punitive objectives against those engaging in statutorily criminalized socially unacceptable behavior while being mentally capable to recognize the unlawfulness of their conduct. Committing a crime always requires a physical element referred to as *actus reus (guilty act).* With the exception of 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* (also known as *culpability*, *fault*, or *blameworthiness*)—criminal law is also plagued by a fair amount of terminological inconsistency. By contrast to tort liability, which takes recourse to mostly objective standards to determine the blameworthiness of a conduct, criminal law measures the defendant's mental state by a predominantly subjective test.

Mens rea encompasses a range of different mental states described by a bewildering variety of terms both within and across legal systems. The three broad categories distinguished are intent (*dolus*), recklessness, and negligence (*culpa*). *Intent* is commonly divided into two—in certain legal systems three—subcategories: (1) *Dolus directus of the first degree* (*direct* or *specific intent*) requires *purposeful* conduct, i.e., that a person commits an offense with the desire to achieve a particular prohibited result. (2) *Dolus directus of the second degree* (*oblique* or *general intent*) is given if an offender acts *knowingly*, that is, intends to commit a prohibited act without desiring to achieve a specific harm but foreseeing its occurrence as virtually certain. The differentiation between these two forms of intent is not always present: for instance Article 30 of the Rome Statute of the International Criminal Court (1998)—a central part of the body of international criminal law—requires the cumulative presence of both volitional and cognitive components. (3) In addition, especially continental criminal legal systems stipulate a third, more indirect notion of intent called *dolus eventualis*, which focuses on the offender's attitude towards the consequences of their action and is satisfied if an individual remains *indifferent* despite foreseeing a possible harm.

Recklessness is an intermediate form of culpable state between intent and negligence in common law jurisdictions, which 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. It is a volitional element without an equivalent in civil law systems, although conscious negligence (explained below) can be regarded as its closest counterpart.

Like in tort law, *negligence* in criminal law also connotes a behavior that departs from the objective standard of conduct of a prudent person. Ordinarily, negligence lacks a cognitive element—which is why English and United States law are divided on whether it counts as a class of mens rea—i.e., the offender *should have been, but was not aware of the substantial and unjustifiable risk* that their action may have negative consequences. Beside this *unconscious negligence*, some jurisdictions distinguish a second form of negligence dubbed *conscious negligence*, given if a person *foresees the risk of harm but believes*—indeed almost hopes—*it will not occur.* To justify the imposition of significantly weightier sanctions, criminal law usually requires *gross negligence*, i.e., considerable deviation from the reasonable person standard. It is fulfillled if an individual's actions pose an *obvious risk* to bring about *substantial harm* and the offender has the *ability to take precautionary measures.* Moreover, negligence is typically only penalized if explicitly criminalized by law—a case in point being Article 30 of the Rome Statute, which excludes criminal responsibility for negligent behavior unless other provisions of the Statute expressly so provide.

Hence, with the exception of strict liability and unconscious negligence offenses, criminal responsibility likewise presupposes that the offender foresees the potential harms their conduct may cause. In those two particular cases, however, only behaviors explicitly criminalized by law entail liability, and such statutory provisions are only conceivable if the legislator foresees that the conduct may result in harm.

In conclusion, we observe that foreseeability—reflecting an inherent ability to control—features 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 economic and political 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 and controllable in the sense required by law.

3. Foreseeability: The Missing Piece of the AI Liability Puzzle

Academic papers, policy documents, and other contributions discussing various aspects of the regulatory treatment of AI typically either handle the concept as given and thus refrain from defining what they mean by AI, or choose a working definition that is best suited to their particular inquiry. While this is not surprising—after all, to date a universally accepted AI definition does not exist—it does create problems of definitional inconsistency. This, in turn, curtails efforts to clearly distinguish between various technologies referred to under the common banner of AI, identify their essence and most relevant properties for distinct policy purposes, and establish an internationally consistent policy stance towards them.

At present, much of the energy dedicated to defining AI is directed to finding some one-size-fits-all definition that is universally applicable in any given context. Insofar as this fosters consistency, we applaud this intent. However, looking at things through a teleological lens highlights that any such broad definition is only of limited use. The purpose of defining AI one way or another is to create a concept with clearly identifiable attributes that we can understand, allowing us to assess AI's capabilities, anticipate its actions and the consequences of those actions, and—ultimately—to make informed decisions on what roles we want it to play in our societies. Yet AI is an umbrella term, which may refer to a number of very different notions from AI as a scientific field (Poole & Mackworth, 2017), to sub-fields of AI like robotics and ML, to specific technologies like differing ML models, the range of which varies over time: Recall, for instance, that what we now know as *big data* was considered AI a few years back. It is apparent that these concepts and/or technologies exhibit distinct characteristics and serve very different purposes, so that they cannot be treated as a single, homogeneous thing.

We are of the opinion that across the manifold use cases and contexts in which we may encounter AI technologies, the technologies themselves—as determined by their state of art at a given point in time—are the only constant elements. The technological characteristics of AI systems determine their capabilities and consequently the tasks a particular system may be used for and the ways in which it may impact human societies. We therefore believe that any AI definition developed for regulatory purposes should first and foremost be based on the particular technology in question. That said, other criteria, like AI use cases and the contexts in which they are deployed, may also be of relevance for constructing specific definitions. Also, even in the presence of such specific definitions, it makes sense to additionally apply a generic definition to provide for consistency within and across domains. The OECD seems to think along similar lines, as demonstrated by their below introduced generic AI system definition laid down in the OECD AI Principles and their upcoming Framework for the Classification of AI Systems, which allows for a technology-specific finetuning of that high-level general AI system definition across four dimensions (Organisation for Economic Co-operation and Development (OECD), 2020).

In this paper we chose to use the term AI system—which we believe is a good example for a generic definition that promotes consistency—as defined in the OECD AI Principles to delineate the group of AI technologies that form the subject of our inquiry. The OECD defines *AI system* as a "machine-based system that" may operate at varying levels of autonomy and "can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments" (Organisation for Economic Co-operation and Development (OECD), 2019). For a non-comprehensive overview of other definitions in circulation see (Council of Europe Ad-Hoc Committee on Artificial Intelligence, 2020a, 2020b).

Our goal in this paper is merely to point out that *there exist* AI systems that present foreseeability issues, rather than to give an exhaustive list of AI technologies that do so—we leave the latter task to further research. In line with that objective, we only consider ML models, a particular class of AI systems. Expanding the taxonomy provided by Flach (2012), we distinguish four main ML models, namely geometric, probabilistic, logic-based, and neural networks (NNs). These four types of model classes form a sort of continuum with logic-based models—which tend to have less foreseeability issues—on one end and NNs—which typically pose the biggest obstacles regarding foreseeability—on the other. This is not to say that logic-based systems are by definition foreseeable while NNs are inevitably impenetrable: Whether and to what extent foreseeability issues arise always depends on the design and complexity of the AI system at hand. But again, here we just want to show that there exist AI systems that may pose foreseeability problems, so instead of going into further technical details, we will only concentrate on NNs to provide such an example.

Owing to low levels of familiarity with AI technologies and the significant hype accompanying AI innovation, most people tend to have an anthropomorphic misconception of AI. This is not only incorrect but also dangerous, as it generates unrealistic expectations of the potential of AI technologies and their impact on society. Therefore, as a preliminary matter, both policymakers and society at large need to be conscious of the fact that AI does not *know*, *think*, *foresee*, *care*, or *behave* in the anthropomorphic sense. Rather, it applies what could be best described as *machine logic*. To illustrate the potential implications of that distinction, consider the following example: 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 (Lipton, 2018). That is, the system identifies outputs based on a set of predefined parameters and probability thresholds through a process that is fundamentally different from human thinking, which typically relies on a mixture of causal understanding, common sense, and emotional impulses.

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 designed at a higher level of abstraction, giving designers less control over how exactly these systems process raw data and identify the right representation they need for solving a particular task. Furthermore, DL-based systems 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 (LeCun, Bengio, & Hinton, 2015).

Importantly, 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 three examples. In the first example, the failure is caused by a *bad classifier* as illustrated by Ribeiro, Singh, and Guestrin (2016) in their Husky vs. Wolf experiment. Here, 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 (or better, white patterns on the lower parts of the pictures) as a classifier for wolves. Thus, in the experiment the system predicted huskies in pictures with snow as wolves and vice-versa. The second and third examples illustrate how *adversarial design*—essentially an automated attack on a search algorithm to minimize its utility—can compromise AI systems: Sharif et al. (2016) designed a *physically executed* attack to cheat a facial recognition system (FRS) by manipulating the physical state the system analyzed rather than its digitized representation. FRS's are usually using NNs in order to recognize patterns in big datasets, in this case the differences between millions of faces (e.g., the relative position of nose and eyebrows, size of the nose, etc.). They used a pair of glasses with a colorful frame, which basically interfered with the system's pattern recognition in two ways: (1) It blocked the *view* to crucial parts of the faces and, (2) due to the colorful frame, gave the system the impression to see some patterns. As a result, the FRS often made mistakes despite indicating a high probability of confidence. Our last example shows a *digitally executed* attack, where the adversary's goal was to create inputs that a DL-based system misclassifies, however, humans do not (Szegedy et al., 2013). 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.

Even without going into technical details, the above analysis shows that it is conceivable that AI systems and the way they generate failures are too complex or otherwise impossible to anticipate and hence do not satisfy the legal foreseeabilty requirement. This insight suggests that we have to develop new legal solutions to attribute liability to such AI systems—or more precisely, given they lack legal personality, their designers, manufacturers, and/or users—as well as to distribute liability between them and their human operator(s). To tackle this challenge, regulators and policymakers will need to dynamically determine which particular AI technologies pose forseeability problems at any given juncture—a task that, in our view, will involve developing and continuously adapting specific definitions tailored to each of those technologies.

4. Adapting the Legal Liability Regime

As shown in the previous sections, holding AI systems—if they acquire legal personality in the future—or their owners, as well as people contributing to their design and/or distribution liable for infringement of legal rights and the resulting damages may not be possible because the failure causing the harm and/or the harm itself were not foreseeable. This leads to considerable uncertainty, which, in turn, significantly hinders AI innovation, as well as trust in and social acceptance of AI technologies. It is a serious problem that needs urgent solution, unless we are willing to accept welfare losses resulting from delaying or even missing out on economic and social benefits AI could bring to humanity. Note that the extent and distribution of aggregate benefits are conditional upon handling AI innovation the right way, especially in a welfare-enhancing rather than economic-inequality-aggravating manner (Korinek & Stiglitz, 2019) and mindful of dual-use concerns (Brundage, M. et al., 2018).

Yet, any reform proposal should carefully balance the objectives a particular field of law seeks to pursue by the imposition of liability and the overarching policy objectives guiding AI innovation. Contract law, where liability for non-performance ultimately aims to achieve an optimal allocation of contractual risks, is the least problematic area, as leaving the matter of contractual risk allocation—through, e.g., negotiation of appropriate guarantee arrangements—to the parties' free disposition will usually yield fair results even with AI's unpredictability. For instance, a seller's willingness to assume risk for an unforeseeable failure of an AI system sold to a buyer can be offset by a higher price negotiated. That said, the framing of policy debates on AI may influence parties' expectations and there may be scope for regulation to correct unjustified fears or overreactions.

As regards tort law, compensation and deterrence are widely recognized as its main policy goals. Loss-spreading, vindication of rights, denunciation of wrongdoing, and educating the public on the proper standards of conduct are mentioned as auxiliary objectives, while views differ on whether private law can or should pursue public interest or have punitive functions (Koziol, 2012; Oliphant, 2015; Green & Cardi, 2015; Posner, 1972). Following in the footsteps of the school of economic analysis of law, many regard civil liability as an instrument to regulate safety (Posner, 1972). This opinion is supported by the fact that tort law falls within the remit of a distinct regulatory strategy recognized in regulatory theory as *allocation of rights and responsibilities* (Baldwin, Cave, & Lodge, 2012). In this view, negligence incentivizes safe conduct up to the economically efficient level (although admittedly without regard to available alternative activities), while strict liability both regulates the socially desired level of hazardous activities and encourages safety (Posner, 1972; Green & Cardi, 2015). An often-heard criticism in the legal community is that such objective economic efficiency analyses are blind to equity considerations and hence not reconcilable with the nuanced legal analysis that is necessary to ensure fair and just outcomes. However, this view seems to ignore that the application of any such analysis requires prior policy choices on which values to maximize and which losses to minimize in order to maximize society's welfare. So while the analyses as decision-making tools are admittedly value-neutral, their use is always preceded by a set of very much value-laden and equity-driven legal and policy judgments (White, 1990). Over time, such judgments have set duties of care for particular negligence torts and designated sources of danger to be addressed by strict liability torts, reflecting a careful and well-established balance of contradicting interests.

Resonating with Hubbard (2016) on the necessity of keeping this delicate balance, we therefore think that it is mistaken to succumb to the temptation to bypass fault or foreseeabilty problems by classifying all instantiations of AI as dangerous and *punishing* their use with the overarching imposition of strict liability. Such an approach would not only be dogmatically incorrect, but would also strangle AI innovation and reduce social welfare compared to an ideal, hypothetical alternative. Given that these outcomes are clearly un-desirable, we suggest a different solution: As mentioned in Section 2, courts in jurisdictions that do not explicitly apply the Hand formula allow social utility considerations to inform their decision as to whether or not a duty of care has been breached in a certain situation. Essentially, this means adding a further variable (U) to the formula, such that $P \times L > B + U$. Relying on this precedent—in instances where the foreseeability requirement is satisfied and hence duties of care can be determined—we propose to expressly account for the economic and social utility derived from AI innovation and it's progressive adoption in society within U. In jurisdictions that currently rely on it in its classic form, this would mean expanding the Hand formula by U. We recognize that doing so would pose some challenges, seeing as it would require development of new methods to quantify, measure, and allocate such utilities to AI innovators, producers, and distributors. Yet, the effort would allow for a more harmonized approach than the currently prevailing practice of leaving these questions up to courts' discretion, which do not necessarily consider economic factors to the desired extent and in any case provide less consistency compared to formal models. More formal calculations of the social utility of an activity would also improve the accuracy of courts' overall balancing exercise in jurisdictions, where a less mathematically explicit approach has been established. Similar considerations should inform the establishment of potential strict liability standards for AI. Given the global relevance of the issue, enhanced consistency and comparability of approaches across jurisdictions would be valuable, not least because it could form the basis of international standards and best practices.

As for criminal liability, hardening liability standards by circumventing the foreseeability requirement is not reconcilable with justifying the imposition of criminal sanctions.

So yes, it is certainly desirable that AI engineers have complete *control* over their systems and avoid design and training failures. But is this a realistic expectation at the present juncture? More importantly, are we willing to stall the adoption of AI until we can *guarantee* its safety? Or is there a compromise that encourages both reasonable safety and innovation?

5. AI Guarantee Schemes as Work-Around

All this leaves us with the problem that, even assuming the legal system will incrementally adapt and solve the above highlighted foreseeability loopholes and other challenges posed by AI, this will take time. As noted by Pearl (2018) in respect of AVs and the US tort system, we are probably looking to several decades of deliberation, trial-and-error type of progress in the legal treatment of AI, and inconsistent jurisprudence. Yet legal certainty is indispensable to get the most out of AI. As outlined in the Introduction, where insurance mechanisms against uncertainty are not available, innovation tends to decrease below socially desired levels, leading to welfare losses (Arrow, 1962). Again, this problem is not specific to AI but common to all new technologies. There have been instances over the course of history, where fears about the legal system's ability to rise to certain challenges have prompted a search for alternative solutions and regulatory interventions on the part of the state.

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 several countries in diverse fields like occupational, medical, and all types of personal injuries. Dispensing with the need to examine how the damage occurred, these systems guarantee victims fast compensation of their claims and involve lower administrative costs compared to litigation.

However, they have the negative effect of promoting carelessness. Moreover, due to financial constraints, they typically only offer partial compensation through the introduction of arbitrary restrictions. Conversely, the judicial system—provided successful litigation—fully compensates victims. 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. Critics of no-fault insurance-based systems, therefore, see little practical difference to litigation and advocate that they complement rather than substitute tort law. See Koziol (2012) for more details.

Another approach, analyzed by Pearl (2018), is setting up *victim compensation funds*—cases in point are 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. Victim compensation funds have been implemented on multiple occasions with varying objectives such as relieving pressure on courts, supporting ailing industries, or simplifying and expediting compensation processes. As regards design questions, 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 the event of litigation. *Non-judicial funds* are divided into three sub-categories: *Public funds*, which are 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*. The latter are also privately administered and funded by private donations, yet they 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 tend to provide flat compensation awards without recourse to tort law to determine eligibility for compensation. One notable advantage of victim compensation funds over conventional litigation is flexibility, given that their status, funding, administration, and processes are customarily designed with a particular set of circumstances in mind. As a general rule, they are also a faster, more efficient, and cost-effective alternative to the tort system. Against these advantages weigh the potentially massive administrative burdens of establishing funds, which by far outweigh their counterparts in the judicial system. Victim compensation funds typically also fall short of providing the degree of transparency and publicity inherent in conventional litigation—attributes that may be of core importance to victims.

Outlining the significant benefits involved, Pearl recommends the creation of a fund for AV crash victims in the United States at least until the legal system catches up with AI innovation. She proposes the establishment of 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. Requiring both buyers and sellers to contribute, she argues, is justified by the fact that 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. Finally, the fund should only cover human injuries and fatalities, whereby 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 any prior insurance payouts.

Our proposal to create a system of *AIGSs* as an insurance mechanism against uncertainty 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 a number of other arrangements—such as 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 (sometimes also referred to as *financial system safety net*) are designed to provide assurance to those involved in financial transactions with financial institutions or markets 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 (Davis, 2004; Schich & Kim, 2011).

Even though perhaps most prominent in the context of financial markets, 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 idea is for the AIGSs to 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. Furthermore, the AIGS should also be available to shore up the legal system in the face of the anticipated uncertainty and complexity of AI-related litigation while courts and policymakers grapple with other novel problems arising from AI. 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. As explained below, there are many open questions regarding the design of the proposed AIGSs. Hence, our goal is once again to spark a high-level, conceptual debate that can inform future policy initiatives in this space, rather than providing a specific example.

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 discussions on AI governance in virtually any domain and country, it is relatively hard to define robust design criteria. Nevertheless—based on Davis' (2004) survey of international practices with respect to FGSs and Pearl's above recommendation on a US national AV victim compensation fund—we will attempt to sketch out an initial set of principles to guide future deliberations on this issue.

Nature of the scheme: Beyond the obvious motivation to provide 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 technologies are beneficial to humanity. One facet of that endeavor is to incentivize AI innovators to employ responsible and safe practices, but the funds could also be instrumental in furthering 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: Among the ranks of academics and various public and private organizations and groupings vested with AI policy development, there is a growing consensus about the necessity of some sort of global governance framework for AI at some point in the future (Erdélyi & Goldsmith, 2018, 2020; Koene et al., 2018). Recent developments in this space include the establishment of several new organizations as well as units or work streams within existing bodies: See for instance the Organisation for Economic Co-operation and Development's AI Policy Observatory (OECD.AI) (Organisation for Economic Co-operation and Development (OECD) AI Policy Observatory, 2020), the Global Partnership on Artificial Intelligence (Global Partnership on Artificial Intelligence, 2020), the World Economic Forum's Centre for the Fourth Industrial Revolution (World Economic Forum (WEF), 2020), the Council of Europe's Ad-Hoc Committee on Artificial Intelligence (Council of Europe, 2020), the Global Governance on AI Roundtable (World Government Summit, 2018), the United Nations Educational, Scientific and Cultural Organization (United Nations Educational, Scientific and Cultural Organization (UNESCO), 2019), the European Commission (European Commission, 2020b, 2020a), and the G7 and G20 groups. However, neither the purview of each of these bodies nor mechanisms for coordination and collaboration between them are clearly defined yet, so they are more of a preliminary institutional architecture than a consistently functioning international governance framework.

As for the domestic level, countries are busy weaving their national AI strategies and passing the most important pieces of legislation to have at least some semblance of control over the most pressing issues across diverse policy domains—like healthcare, financial services, the criminal justice system, or welfare—without much regard to cross-sectoral consistency.

Hence, it would seem that the reality is still that AI innovation and implementation is outpacing policymakers' regulatory and oversight capabilities. As the 2018 AI Now Report (Crawford et al., 2018) acknowledges, each of these distinct domains has its established regulatory frameworks, traditions, and specific difficulties, requiring specialized expertise and sector-specific regulation. 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 in the interim, until more research can be done on the optimality of governance arrangements. In the context of FSG governance, the Davis report (2004) 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) keep the administrative costs of the fund as well as (4) compliance burdens for industry as low as possible, (5) where appropriate, involve industry stakeholders, harnessing their expertise, and (6) provide an adequate incentive structure for regulatory authorities.

Taken together, these observations furnish strong arguments to house AIGSs within domain-specific agencies, at least until experience provides us with more clarity on the vices and virtues of such an approach. 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 and recursive whole-of-society dialogue. Once up and running, such cross-jurisdictional governance arrangements could justify a transnationally organized AIGS system—potentially in addition to and complementing domestic systems. Although a glance at financial regulatory experience gives reason to doubt the practical feasibility of any sort of global plans and we realize that the prospect is in any case a remote one, it should still not be entirely taken off the table.

Coverage: As noted by the Davis report (2004), fund coverage design inevitably involves wrestling and eventually putting up with tradeoffs between the conflicting objectives of efficiency, equity, and minimum complexity and cost. Note that 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., of the establishment, ongoing operation of schemes, and 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 widespread opinion, in view of guarantee schemes' role as safety net—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. Nevertheless, few FGSs leave this matter to financial institutions' discretion. Instead, they 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 in fact be preferable. Such an approach could additionally be justified by AIGSs intended rational as a tool to regulate the AI industry's incentive structure, while at the same time potentially pursuing other policy objectives.

Funding and pricing: Guarantee schemes involve a redistribution of losses, calling certain stakeholders to foot the bill to alleviate pressure on others. Striking a level of redistribution that stakeholders perceive as fair is therefore key to ensure guarantee funds' acceptance and efficiency. Funding relates to the timing and rate of contributions, as well as the base of funding, while pricing determines contributors' relative share. With respect to FGSs, the Davis report (2004) notes that funding and pricing considerations should strive to accommodate four general goals: (1) cost efficiency (minimize administrative costs), (2) competitive neutrality (equitable treatment of contributors with 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, fund administrators have the choice between (1) pre-funding, where contributions are paid into and managed by the fund, (2) post-funding, whereby contributors incur contingent liabilities and are only required to pay into the fund after the guarantee triggering event, or (3) a combination of both. Pre-funding usually implies greater stability and credibility that funds are readily available in the event of a 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.

Regarding the funding base, the main 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. However, 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.

Finally, pricing choices are usually about striking an acceptable balance between simplicity and efficiency. The latter is promoted by differential, risk-sensitive contributions, which are typically the better choice when it comes to combating moral hazard and ensuring equitable treatment of contributors, but are also 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, the kinds of systemic crises with the potential to deplete FGSs and necessitate state involvement are admittedly a highly remote possibility in the AI context. However, since it is impossible to predict the exact trajectory of AI innovation, it is hard to anticipate if and how this might change in the future. This uncertainty coupled with better feasibility of risk-sensitive pricing and the likelihood that industry would perceive pre-funding as the fairer funding option are strong arguments in favor of pre-funding. Because of the lack of large-scale shocks that may strain schemes' funding resources, it is unlikely that post-funding, even in an auxiliary form, will be necessary—again, this may change based on how the current state of affairs develop. As for funding base, we believe that, unless specific policy considerations dictate otherwise, this 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. Reiterating the importance of the perceived fairness of schemes' redistributive effects, we strongly favor risk-sensitive pricing arrangements. We also call for considering a number of risk management techniques available in 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.

We would like to conclude our proposal on establishing a system of AIGSs with two final thoughts: (1) Whether an innovation is Pareto-improving, crucially depends on how its revenues/rents are distributed across society. While redistribution is costly, entailing a tradeoff between equity and efficiency, this tradeoff does not occur if regulatory frameworks for innovations are designed with certain distributive goals in mind from the outset (Korinek & Stiglitz, 2019). (2) As noted earlier, the economy is structured by a complex regulatory framework, which liability rules are part of. Hence, prior to any regulatory intervention, it must be considered how this would affect the existing status quo and what combination of rules changes is most conducive to achieve the set regulatory goals.

The corollary of this in the AI context is that we cannot look at isolated AI liability or guarantee scheme rules, but must treat the economy and the structuring regulatory framework as a system. Also, in this initial stage of developing AI regulatory frameworks, we have an opportunity to design these in a holistic manner, avoiding inequality aggravating effects and inefficiencies resulting from the necessity of later redistributive adjustments.

6. 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 put forward solutions that would facilitate a smooth transition into an AI-driven society.

One aim was to expand the existing literature by providing a comparative legal analysis spanning both civil and criminal legal domains to make the claim that foreseeability is a central prerequisite for attributing legal liability across all jurisdictions and legal domains. We then showed that there exist certain AI systems, which do not satisfy the foreseeability requirement, making it impossible to solve liability issues via conventional legal liability regimes in some circumstances, and generating considerable legal uncertainty. We also raised economic and international relations arguments to highlight the economic and political costs of treating liability problems as a solely legal matter and of failure to resolve this problem of uncertainty in a timely manner. To assist current policy efforts to settle on a widely-accepted AI definition, we engaged in a discussion of the vices and virtues of such an approach, advocating for functional, technology-specific definitions for regulatory and policy purposes with an auxiliary role for more generic definitions.

The recommended amendment to the legal liability system would better account for AI's social utility. The system of AIGSs—an insurance mechanism against uncertainty outside of the purview of legal liability—would constitute a predictable and transparent framework for swift compensation of damages where litigation is not promising either because the category of harm caused by an AI system is unforeseeable and hence not imputable under current legal liability rules or because the process would be overly complex due to other legal intricacies. Prospective defendants would no longer need to fear arbitrary court decisions that stretch the limits of legal liability in dogmatically inconsistent, unpredictable ways to correct an otherwise uncompensated injustice. Potential victims and aggrieved parties would have peace of mind using AI, knowing that bringing complex and expensive actions of dubious outcome are no longer the only option to obtain compensation should something go wrong. By fair loss-spreading and clear allocation of risks among potential defendants and plaintiffs or prosecutors in future AI liability suits, the proposed system of AIGSs would also support emerging markets in AI technologies, in particular foster innovation and AI's social acceptance. Moreover, if desired, the AIGSs could assume a broader role within the overall regulatory framework structuring our economies. Of course, the current virtually non-existent AI governance landscape leaves quite a few blanks in regard to the AIGSs' design, but reassuringly, experience with FGSs could inform many design and implementation decisions AIGS designers are likely to face. In sum, drawing on best-practice mechanisms in financial regulation, AIGSs would provide legal certainty in dealing with AI-related liability issues without violating existing liability doctrines and induce a legal environment that fosters safe and responsible AI innovation and adoption in society.

In line with our objective to point out and raise awareness towards a general, conceptual legal problem and also because the scope of the present paper did not allow for addressing the relevant case law of multiple jurisdictions, we expressly leave this work for future research.

*A*cknowledgments: We are grateful to the anonymous reviewers of the Journal of Artificial Intelligence Research for their extremely helpful, thorough, and constructive comments and literature pointers.

References

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

Arrow, K. (1962). Economic Welfare and the Allocation of Resources for Invention. In Universities-National Bureau Committee for Economic Research and Committee on Economic Growth of the Social Science Research Council (Ed.), *The Rate and Direction of Inventive Activity: Economic and Social Factors*, pp. 609–626. Princeton University Press.

Baldwin, R., Cave, M., & Lodge, M. (2012). *Understanding Regulation: Theory, Strategy, and Practice* (2 edition). Oxford University Press.

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

Brundage, M. et al. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. Tech. rep. 1802.07228, CoRR.

Brunner, C. (2009). *Force Majeure and Hardship under General Contract Principles* (1 edition). Kluver Law International.

Bryson, J., Diamantis, M., & Grant, T. D. (2017). Of, for, and by the People: The Legal Lacuna of Synthetic Persons. *Artificial Intelligence and Law*, *25*, 273–291.

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

Council of Europe (2020). Ad-Hoc Committee on Artificial Intelligence (CAHAI). https://www.coe.int/en/web/artificial-intelligence/cahai.

Council of Europe Ad-Hoc Committee on Artificial Intelligence (2020a). Feasibility Study. https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da.

Council of Europe Ad-Hoc Committee on Artificial Intelligence (2020b). Towards Regulation of AI Systems. https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a.

Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Sánchez, A. N., Raji, D., Rankin, J. L., Richardson, R., Schultz, J., West, S. M., & Whittaker, M. (2018). AI Now Report. https://ainowinstitute.org/AI_Now_2019_Report.pdf. AI Now 2019 Report, New York, AI Now Institute.

Davis, K. (2004). Study of Financial System Guarantees. https://treasury.gov.au/publication/p2004-45061. Davis Report.

Erdélyi, O. J., & Goldsmith, J. (2018). Regulating Artificial Intelligence: Proposal for a Global Solution. In *Proceedings of the First AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, pp. 95–101.

Erdélyi, O. J., & Goldsmith, J. (2020). Regulating Artificial Intelligence: Proposal for a Global Solution. Tech. rep. 2005.11072, arXiv.

European Commission (2020a). AI Watch. https://ec.europa.eu/knowledge4policy/ai-watch_en.

European Commission (2020b). Shaping Europe's digital future: Artificial Intelligence. https://ec.europa.eu/digital-single-market/en/artificial-intelligence.

European Commission (2020c). White paper on artificial intelligence: A european approach to excellence and trust. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

European Commission High-Level Expert Group on Artificial Intelligence (2019). Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Flach, P. (2012). *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press.

Fletcher, G. P. (2000). *Rethinking Criminal Law* (1 edition). Oxford University Press.

G20 (2019). G20 Ministerial Statement on Trade and Digital Economy. http://www.g20.utoronto.ca/2019/2019-g20-trade.html. Meeting of 8 and 9 June 2019.

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

Global Partnership on Artificial Intelligence (2020) https://gpai.ai.

Green, M. D., & Cardi, W. J. (2015). Basic Questions of Tort Law from the Perspective of the USA. In Koziol, H., & Askeland, B. (Eds.), *Basic Questions of Tort Law from a Comparative Perspective*, pp. 431–514. Jan Sramek Verlag.

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

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

Institute of Electrical and Electronics Engineers (2019). Ethically Aligned Design, First Edition. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf?utm_medium=undefined&utm_source=undefined&utm_campaign=undefined&utm_content=undefined&utm_term=undefined.

International Criminal Court (1998). Rome Statute. https://www.icc-cpi.int/nr/rdonlyres/ea9aeff7-5752-4f84-be94-0a655eb30e16/0/rome_statute_english.pdf.

Judge Learned Hand (1947). United States v. Carroll Towing Co., 159 F.2d 169 (2d Cir. 1947). Hand Formula.

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

Karnow, C. E. A. (2016). The Application of Traditional Tort Theory to Embodied Machine Intelligence. In Calo, R., Froomkin, A. M., & Kerr, I. (Eds.), *Robot Law*, pp. 51–77. Edward Elgar Publishing.

Kingston, J. K. C. (2016). Artificial Intelligence and Legal Liability. In Bramer, M., & Petridis, M. (Eds.), *Research and Development in Intelligent Systems XXXIII*, pp. 269–279. Springer International Publishing.

Koene, A., Clifton, C., Hatada, Y., Webb, H., Patel, M., Machado, C., LaViolette, J., Richardson, R., & Reisman, D. (2018). A Governance Framework for Algorithmic Accountability and Transparency.. EPRS/2018/STOA/SER/18/002 European Parliament Science Technology Options Assessment report presented at the European Parliament on October 25th 2018.

Köhler, S. (2018). Instrumental robots. Science and Engineering Ethics https://doi.org/10.1007/s11948-020-00259-5.

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

Koziol, H. (2012). *Basic Questions of Tort Law from a Germanic Perspective* (1 edition). Atlasbooks Dist Serv.

Koziol, H., & Askeland, B. (2015). *Basic Questions of Tort Law from a Comparative Perspective* (1 edition). Jan Sramek Verlag.

LeCun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Liechtung, J. (2018). The Race is On! Regulating Self-Driving Vehicles Before They Hit The Streets. *Brooklyn Journal of Corporate, Financial & Commercial Law*, *12*(2), 389–413.

Lipton, Z. C. (2018). The Mythos of Model Interpretability. *Queue*, *16*(3), 30:31–30:57.

Marchuk, I. (2014). *The Fundamental Concept of Crime in International Criminal Law: A Comparative Law Analysis* (1 edition). Springer Berlin Heidelberg.

National Transportation Safety Board (NTSB) (2018). Highway Accident Report: Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian. https://data.ntsb.gov/Docket?NTSBNumber=HWY18MH010.

Oliphant, K. (2015). Basic Questions of Tort Law from the Perspective of England and the Commonwealth. In Koziol, H., & Askeland, B. (Eds.), *Basic Questions of Tort Law from a Comparative Perspective*, pp. 355–430. Jan Sramek Verlag.

Organisation for Economic Co-operation and Development (OECD) (2019). Recommendation of the Council on Artificial Intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. Adopted on 22 May 2019, C/MIN(2019)3/FINAL.

Organisation for Economic Co-operation and Development (OECD) (2020). A First Look at the OECD's Framework for the Classification of AI Systems, Designed to Give Policymakers Clarity. https://oecd.ai/wonk/a-first-look-at-the-oecds-framework-for-the-classification-of-ai-systems-for-policymakers.

Organisation for Economic Co-operation and Development (OECD) AI Policy Observatory (2020). https://oecd.ai..

Pearl, T. (2018). Compensation at the Crossroads: Autonomous Vehicles and Alternative Victim Compensation Schemes. In *Proceedings of the 29th European Regional Conference of the International Telecommunications Society (ITS): "Towards a digital future: Turning technology into markets?"*.

Perc, M., Ozer, M., & Hojnik, J. (2019). Social and Juristic Challenges of Artificial Intelligence. *Palgrave Communications*, *5*(61), 1–7.

Poole, D., & Mackworth, A. (2017). *Artificial Intelligence: Foundations of Computational Agents* (2 edition). Cambridge University Press, Cambridge, UK.

Posner, R. A. (1972). A Theory of Negligence. *The Journal of Legal Studies*, *1*(1), 29–96.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM.

Schich, S., & Kim, B.-H. (2011). Guarantee Arrangements for Financial Promises: How Widely Should the Safety Net be Cast?. *OECD Journal: Financial Market Trends*, *2011*(1), 201–235.

Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2016). Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1528–1540. ACM.

Stiglitz, J. E. (2015). Rewriting the Rules of the American Economy: An Agenda for Growth and Shared Prosperity. https://rooseveltinstitute.org/wp-content/uploads/2015/05/RI-Rewriting-the-Rules-201505.pdf. Roosevelt Institute.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., & Fergus, R. (2013). Intriguing Properties of Neural Networks. Tech. rep. 1312.6199, CoRR.

Templeton, B. (2019). NTSB Hearing Blames Humans, Software and Policy for Fatal Uber Robocar Crash — But Mostly Humans. https://www.forbes.com/sites/bradtempleton/2019/11/19/ntsb-hearing-blames-humans-software-and-policy-for-fatal-uber-robocar-crash/?sh=5dfd4bc64c6d.

United Nations Educational, Scientific and Cultural Organization (UNESCO) (2019). https://en.unesco.org/artificial-intelligence..

United Nations (UN) (1980). UN Convention on Contracts for the International Sale of Goods. https://www.uncitral.org/pdf/english/texts/sales/cisg/V1056997-CISG-e-book.pdf.

Vladeck, D. C. (2014). Machines Without Principals: Liability Rules and Artificial Intelligence. *Washington Law Review*, *89*(117), 117–150.

Weinrib, E. J. (2012). *The Idea of Private Law* (1 edition). Oxford University Press.

White, B. A. (1990). Risk-Utility Analysis and the Learned Hand Formula: A Hand That Helps or a Hand That Hides?. *Arizona Law Review*, *32*(1), 77–136.

World Economic Forum (WEF) (2020). Centre for the Fourth Industrial Revolution. https://www.weforum.org/centre-for-the-fourth-industrial-revolution/home.

World Government Summit (2018). Global Governance on AI Roundtable (GGAR). https://ggar.worldgovernmentsummit.org/en.