Criminal Law Forum (2023) 34:43–85
https://doi.org/10.1007/s10609-023-09451-1

ALICE GIANNINI* and JONATHAN KWIK

NEGLIGENCE FAILURES AND NEGLIGENCE FIXES. A COMPARATIVE ANALYSIS OF CRIMINAL REGULATION OF AI AND AUTONOMOUS VEHICLES

Accepted: 28 December 2022; Published online: 12 January 2023

ABSTRACT. Automated vehicles (“AV”) can greatly improve road safety and societal welfare, but legal systems have struggled with the prospect of whom to hold criminally liable for resulting harm, and how. This difficulty is derived from the characteristics of modern artificial intelligence (“AI”) used in AV technology. Singapore, France and the UK have pioneered legal models tailored to address criminal liability for AI misbehaviour. In this article, we analyse the three models comparatively both to determine their individual merits and to draw lessons from to inform future legislative efforts. We first examine the roots of the problem by analysing the characteristics of modern AI vis-à-vis basic legal foundations underlying criminal liability. We identify several problems, such as the epistemic problem, a lack of control, the issue of generic risk, and the problem of many hands, which discommode the building blocks of criminal negligence such as awareness, foreseeability and risk taking – a condition we refer to as negligence failures. Subsequently, we analyse the three models on their ability to address these issues. We find diverging philosophies as to where to place the central weight of criminal liability, but nevertheless identify common themes such as drawing bright-lines between liability and immunity, and the introduction of novel vocabulary necessary to navigate the new legal landscape sculpted by AI. We end with specific recommendations for future legislation, such as the importance of implementing an AI training and licensing regime for users, and that transition demands must be empirically tested to allow de facto control.

I INTRODUCTION

Criminal law has always had a strenuous relationship with automatons, with each new generation of automation triggering discussions on whether their introduction into society would generate problems

* Alice Giannini is a Joint PhD Candidate in Criminal law at the Faculty of Law, University of Florence, Italy, and at the Faculty of Law, University of Maastricht, Netherlands. E-mail: a.giannini@maastrichtuniversity.nl. Jonathan Kwik is a PhD Candidate in Criminal law at the Faculty of Law, University of Amsterdam, Netherlands. E-mail: h.c.j.kwik@uva.nl.

Sections I and IV were written jointly by the authors. Section III, including its subsections, were written by Alice Giannini. Section II, including its subsections, were written by Jonathan Kwik.

or loopholes in the field of criminal law. Controversies related to the liability for acts of automatons can be traced as far back as the early 19th century.[1] In that sense, modern strands of Artificial Intelligence (''AI'') merely comprise the latest iteration of this debate. However, modern AI, characterised by machine learning (''ML'') and dynamic decision-making, constitutes a substantial step forward when compared to older forms of automation, even when contrasted to the previous generation of rule-based AI. Perri 6, one of the first seminal authors to have written about this topic, predicted in 2001 that once a machine achieved a certain level of autonomy, difficulties would arise in attributing responsibility.[2] A few years later, Matthias first coined the term ''responsibility gap'', a phrase we encounter often in discussions around AI today.[3] Matthias specifically identified the lack of predictability and control in modern AI as the main obstacle in identifying a responsible subject. Attributing blame is indeed difficult if neither the designer nor the operator can predict how an AI will react, as it learns from experience and acts in accordance with its environment.

Since these early years, the sophistication of our AI technologies has evolved significantly, and so has our experience with these so-called responsibility gaps. As AI is being deployed in increasingly high-risk domains such as driving, it is thus unsurprising that legislatures have begun to introduce measures to ensure the equitable administration of justice for resultant harms.[4] These measures, as it will be shown, rely heavily on attaching criminal liability to a human's capacity of foreseeing and handling risks, ie, to whether they were (criminally) negligent. In this regard, when discussing issues of AI and *mens rea*, specifically those related to *culpa* and autonomous vehicles (''AVs''), we will introduce the concept of ''negligence failures''.

---

[1] Ugo Pagallo, ''From Automation to Autonomous Systems: A Legal Phenomenology with Problems of Accountability'', *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (2017) p. 17.

[2] Perri 6, ''Ethics, Regulation and the New Artificial Intelligence, Part II: Autonomy and Liability'' (2001) 4 *Information, Communication and Society*.

[3] Andreas Matthias, ''The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata'' (2004) 6 *Ethics and Information Technology*.

[4] Gabriel Hallevy, ''Unmanned Vehicles – Subordination to Criminal Law Under the Modern Concept of Criminal Liability'' (2012) 21 *Journal of Law, Information and Science*, 200.

Negligence failures can be defined as situations in which the classical building blocks of negligence, ie, risk taking, foreseeability, and awareness, struggle to identify a liable human being to whom we can attribute AI-caused harm. One could envisage negligence failures as nothing but a further development of the ''irreducibility challenge'', first theorized by Abbott and Sarch,[5] applied specifically in the field of criminal negligence. We also argue that the crude fix of requiring permanent human oversight – as proposed by some parties – clashes with cognitive perspectives and criminal legal theory, and often nullifies the advantages automation is intended to provide in the first place. Indeed, more refined regimes or interpretations of the law are necessary to avoid unequitable attribution of responsibility or scapegoating.

Recently, three countries have addressed the matter: Singapore, France and the UK. We will analyze their approaches in this order, as the Singaporean proposal provides an idea of a *general* framework of criminal liability connected to AI systems, while the French and the UK models provide an example of two different *sector specific* frameworks of criminal liability connected to AI systems, ie, autonomous driving. These three countries address negligence failures through novel legal constructs, such as the creation of immunity clauses, new legal subjects, such as the ''user-in-charge'', and specific criminal offences for producers in cases where users are misled as to the AI system's functioning. The proposals will be scrutinized from the perspective of criminal law, touching upon their efficacy in addressing problems caused by the introduction of AI decision-making for high-risk tasks (such as driving), by critically examining their advantages and disadvantages. We will closely consider whether the legal constructs (ie, the ''fixes'') they propose are in line with principles of criminal law in general and *mens rea* requirements in particular. Moreover, we will identify specific shortcomings of these fixes, for example the failure to properly consider issues specific to modern AI such as bias, data dependency, and the fact that an AI producer (or ''programmer'') is not a monolithic entity.

This article is structured as follows. First, in Section II, we provide context to our discussion by examining in greater detail the main problem the Singaporean, French and British proposals are meant to solve, ie, the roots of negligence failures. These primarily relate to the ''epistemic problem'' and the ''control problem'', although we also

[5] Ryan Abbott and Alex Sarch, ''Punishing Artificial Intelligence: Legal Fiction or Science Fiction'' (2019), 53 *UC Davis Law Review*.

address related issues such as generic risk and the problem of many hands. Subsequently, in Section III, we outline the Singaporean, French, and British approaches respectively. We then identify and evaluate both similarities and dissimilarities amongst the three approaches. Finally, we conclude with a summary of our findings and recommendations for future legal discussions, developments, and policy initiatives.

II AI AND THE STRUGGLES OF CRIMINAL LIABILITY

To properly assess the efficacity of the proposals discussed in Section III of this paper, it is useful to first obtain a solid understanding of what they are meant to "fix". Therefore, in this section we will highlight exactly the factors which make modern AI problematic in terms of fair allocation of criminal liability, ie, what negligence failures imply.

Before proceeding, it may be useful to briefly outline what constitutes "modern AI" and their near-future prospects. The modern paradigm of AI can be characterised mainly by the ubiquity of machine learning techniques.[6] The use of deep neural networks, increasingly powerful GPUs and the availability of collecting massive databases considerably improved their performance and possible applications.[7] For many tasks, AI is found to consistently outperform humans,[8] providing strong incentives for both States and the private sector to invest in its development and use.[9] As Lohn remarks, "AI is

[6] House of Lords, "Select Committee on Artificial Intelligence, Report of Session 2017–2019, AI in the UK: Ready, Willing, and Able?" (2018) HL Paper 100, 16 April 2018, p. 19.

[7] A. Vogelsang and M. Borg, "Requirements Engineering for Machine Learning: Perspectives from Data Scientists" (2019) < http://arxiv.org/abs/1908.04674 >, p. 1.

[8] Yampolskiy provides an impressive list of AI accomplishments in the period of 2004–2016, which also illustrates the exponential development in AI performance in a relatively short amount of time. See R.V. Yampolskiy and M.S. Spellchecker, "Artificial Intelligence Safety and Cybersecurity: A Timeline of AI Failures" (2016) <http://arxiv.org/abs/1610.07997>, pp. 1–2.

[9] C. Gao, "China Vows to Become an Artificial Intelligence World Leader" (*The Diplomat*, 21 July 2017) <https://thediplomat.com/2017/07/china-vows-to-become-an-artificial-intelligence-world-leader> accessed 25 October 2020 (China); Anna Roy, *National Strategy for Artificial Intelligence #AIFORALL* (National Institution for Transforming India Aayog 2018) 5 (India); Joint Research Centre and AI Watch, *Defining Artificial Intelligence: Towards an operational definition and taxonomy of artificial intelligence* (European Union 2020) 6 (Europe).

a ubiquitous technology that can be envisioned in an infinity of applications.'' [10] Currently, AI is employed even in high-risk contexts, such as the medical sector, navigation (autopilots, AVs), loan rejection, recidivism prediction, flight risk prediction for bails, and weapons.[11]

However, incorporation of AI can also be dangerous. While it is projected that AI performance will continually improve in the coming years,[12] there is little prospect of exponential leaps in the near future.[13] Even in very optimistic projections, 2040 is regarded as the earliest moment artificial general intelligence (AGI) can even be considered a possibility.[14] In the near future, therefore, AI will remain ''narrow'': ie, their high level of performance can only be maintained in ''one or few specific tasks''.[15] Even within the tasks they are designed to perform, 100% reliability is impossible to

[10] A.J. Lohn, ''Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing Out-Of-Distribution Performance'' (2020) <http://arxiv.org/abs/2009.00802>, p. 2.

[11] See ia: A. Abdul *et al*., ''Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda'' (2018) *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 5; P. Scharre, *Army of None* (W.W. Norton, 2018) 192; G. Bansal, ''Explanatory Dialogs: Towards Actionable, Interactive Explanations'' (2018) *2018 AAAI/ACM Conference*, p. 356.

[12] Roy, *supra* note 9, 13. In contrast, exact numbers are difficult to provide and vary per domain, as it is very hard to predict future technological innovations that could lead to sudden and drastic improvements in processing power, sensor sophistication, etc. See e.g. Scharre, *supra* note 11, 347. As such, in this article, we will not base our analysis on these more fluctuating variables (e.g. quantitative performance projections) but instead on more stable assumptions related to fundamental ML characteristics (e.g. brittleness, opacity) which are likely to remain applicable in the foreseeable future as long no paradigm shift occurs away from today's ML-centric AI.

[13] J.G. Thorne, *Warriors and War Algorithms: Leveraging Artificial Intelligence to Enable Ethical Targeting* (Naval War College, 2020) 7; S.J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach* (3rd edn, Pearson 2010), 1020.

[14] V.C. Müller and N. Bostrom, ''Future Progress in Artificial Intelligence: A Survey of Expert Opinion'' *in* V.C. Müller (ed), *Fundamental Issues of Artificial Intelligence* (Springer 2016).

[15] Independent High-Level Expert Group on Artificial Intelligence (HLEG), *A definition of AI: Main capabilities and disciplines* (European Commission 2019) 5.

achieve: "An AI designed to do X will eventually fail to do X." [16] If employed in high-risk situations, they may fail spectacularly and lead to significant harm or damage. Additionally, challenges related to (lack of) understandability,[17] bias,[18] and attributing accountability[19] have raised concern. Notwithstanding these limitations, humanity seems to have accepted AI as a way to improve efficiency and societal welfare. This does not mean, however, that these concerns should be ignored, and efforts are being made in both the technical and policy domain to discuss and address these challenges. The current article focuses on one important aspect: criminal responsibility.

There are several characteristics of modern AI systems which make fulfilling all requirements for criminal liability challenging. A great majority of these concerns relate to the *mens rea* component. While some authors have rightly pointed out that the *actus reus* requirement can potentially also be problematic to establish,[20] we will focus primarily on the *mens rea* component in this section. *Mens rea* generally requires knowledge and volition, and it is particularly this knowledge that is called into question with regard to complex systems based on ML or hybrid architectures.

This being said, some AI-related circumstances are *not* conceptually problematic, hence we will not address them further in our analysis. These are cases where there is intent to commit a crime *using* an AI system. Evidently, like any other tool, AI can be deliberately misused by nefarious actors for criminal purposes. Real-world examples of such scenarios are plentiful, such as theft, financial fraud, forgery, market manipulation, phishing, deepfakes and cyberattacks.[21] Hayward et al. developed a useful typology in this regard,

[16] Yampolskiy and Spellchecker, *supra* note 8, p. 5.

[17] See Section 2.2.

[18] "Bias" here refers to the situation where the ML system integrates undesirable patterns of bias (e.g. ethnic discrimination) during its training, and therefore exhibits the same bias during operation. See generally A. Weller, "Transparency: Motivations and Challenges" (2019) < http://arxiv.org/abs/1708.01870 >; D. Danks and A. J. London, "Algorithmic Bias in Autonomous Systems" (2017) *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. Note that this must be distinguished from the term "automation bias" used in a latter section of this article. Cf Section 3.3.

[19] See Matthias, *supra* note 3; Pagallo, *supra* note 1.

[20] T.C. King and others, "Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions" (2019) 26 *Science and Engineering Ethics*, 95.

[21] *ibid* 90–106.

distinguishing between crimes *with* AI (as a tool), *on* AI (as an attack surface) and *by* AI (as an intermediary).[22] Similarly, an AV could hypothetically be intentionally programmed to ram into specific ethnic groups on the sidewalk or be activated deliberately by a driver in unsuitable conditions to provoke a collision. These cases are egregious but less conceptually troublesome, as criminal law is generally well-equipped to handle instances of deliberate acts: "Criminal culpability is self-evident in the case of intent." [23] In such situations, the autonomy or sophistication of the AI is less relevant, as it would constitute "nothing but a tool in the criminal hands of human agents",[24] thus engendering their criminal liability.[25] While it is true that prosecutors may still encounter *evidentiary* obstacles in proving such intent,[26] there is nothing about modern AI that produces an *inherent* responsibility gap in such scenarios. As we will see below, it is the cases where deliberate intent[27] is lacking which are truly challenging.

2.1 *Risk-taking in Terms of Mens Rea*

If deliberate intent is lacking, then the accused might have engaged in (culpable) risk-taking. Unfortunately, there is no common (legal) language on the typification of the different kinds of *mens rea* in this respect. While roughly every modern legal system recognizes intention or purpose (*dolus directus*), categorizations differ when it comes to the remaining forms of guilty mental states.[28] For example,

[22] K.J. Hayward and M.M. Maas, "Artificial Intelligence and Crime: A Primer for Criminologists" (2021) 17 *Crime, Media, Culture: An International Journal*.

[23] Nikolas Stürchler and Michael Siegrist, "A 'Compliance-Based' Approach to Autonomous Weapon Systems" (*EJIL Talk*, 2017) <www.ejiltalk.org/a-compliance-based-approach-to-autonomous-weapon-systems> accessed 7 June 2021.

[24] Daniele Amoroso and Benedetta Giordano, "Who Is to Blame for Autonomous Weapons Systems' Misdoings?" *in* Elena Carpanelli and Nicole Lazzerini (eds), *Use and Misuse of New Technologies* (Springer International Publishing 2019), p. 217.

[25] King, *supra* note 20, 109.

[26] See also below, Subsection 2.5.

[27] In the sense of *dolus directus*.

[28] For example the Model Penal Code at § 2.02 (2)(a) distinguishes between intention (conscious desire to bring about the result); knowledge (of the forbidden result which will almost certainly follow the act); recklessness (doing an act realizing that it involves a substantial and unjustifiable risk of harm); and negligence (objective fault in creating an unreasonable risk). Wayne LaFave, *Modern criminal law* (West 1988), p. 243.

according to the classification provided in the Model Penal Code, the difference between *recklessness* and *negligence* is not the risk created, which is the same (ie, a substantial and unjustifiable risk), but rather the fact that first entails that the agent "is aware that her conduct creates a substantial and unjustifiable risk",[29] whereas the second does not. In other words, according to this model the reckless actor *consciously* disregards the risk, while the negligent actor does not.

Most continental legal systems, instead, are based on a bipartite scheme of (guilty) mental states, which includes only intent (*dolus*) and negligence (*culpa*). The latter, then, encloses the other "intermediate modes"[30] of subjective responsibility, such as recklessness. In these systems, scholars and jurisprudence struggle to identify where to place the conduct of "conscious risk taking".[31] The *escamotage* is to be found in the *dolus eventualis* and *conscious negligence* doctrines. *Dolus eventualis* can be described as a conduct of *intentional risk taking*: "the actor does not know whether his conduct will bring about a harmful result but accepts the occurrence of that result 'in the event that' it comes about".[32] In other words, the agent "mentally embraces that outcome".[33] *Conscious negligence*, instead, can be described as a conduct of *negligent risk taking*: the actors do not know whether their conduct will bring about a harmful result, in fact, they unreasonably reject the idea or do not take this possibility seriously[34] (a sort of "everything will be alright" kind of mental state),[35] but still decide to take the risk.[36]

Having acknowledged this, it is not necessary for the purposes of the current discussion to adopt a stance in this debate. What is more

[29] Luis E. Chiesa, "Mens Rea in a Comparative Perspective" (2018) 102 *Marquette Law Review*, 581.

[30] Thomas Weigend, "Subjective Elements of Criminal Liability", in Markus D. Dubber and Tatjana Hörnle (eds), *The Oxford Handbook of Criminal Law* (OUP, 2015) p. 498. See also George Fletcher, "The Theory of Criminal Negligence: a Comparative Analysis" (1971), 119 *University of Pennsylvania Law Review*, 401.

[31] Weigend, *supra* note 30, p. 500.

[32] *ibid*

[33] *ibid*

[34] *ibid*

[35] *ibid*

[36] Some argue that the distinguishing element between the two should not be the volition element, rather the knowledge one: the real difference between *dolus eventualis* and conscious negligence (or *luxuria*), then, lies in whether the agent knew that there was a *grave* risk of harm or a *minor* risk of harm. *ibid* p. 501.

important to retain is that some forms of criminal liability are based on the fact that an agent had – more or less strongly – *foreseen* and – more or less strongly – *accepted the risk* that an unlawful consequence will arise from the conduct.[37]

Finally, an aspect of negligence with is relevant for this inquiry regards the specificity of one's foresight which is needed to establish negligence. Such an evaluation presupposes understanding whether the agent should have foreseen the *specific* harmful consequence which resulted from their conduct, eg, the specific dynamic of a car accident, or whether, instead, it is sufficient that the agent foresaw a general risk of harm. Common law scholars, and courts, refer to this evaluation as the "reasonable foreseeability test".[38] If we decline this to an AV-scenario, we might ask ourselves to what extent the (unpredictable) functioning of such systems could pose as an "unreasonable" source of harm, which could prove ungovernable for the driver, hence leading to exemption from liability. This issue is particularly important in areas which always involve "some" risk, such as driving a car on a public road, as it will be discussed in Section 2.3.

2.2 *The Epistemic Problem*

Let us now apply this framework to a more concrete situation concerning AVs, say a crash involving pedestrians. According to the *mens rea* theory outlined above, to hold a driver criminally liable for activating their autonomous vehicle, which subsequently killed a family down the hill, this person must have been able to foresee this result – or at least the risk that it could manifest. How such a consequence would manifest should be clear for the person who starts driving with defective breaks, but not necessarily so for the owner of an AV which glitched momentarily because of a reflection off a rooftop. Himmelreich refers to this matter as the *epistemic problem*, ie, difficulties in establishing responsibility because the accused lacks the necessary foresight, foreseeability, or awareness.[39]

[37] *ibid* p. 498.

[38] The test can be summarized as follows "did the defendant have to reasonably foresee the 'exact form' of subsequent act, or was foreseeability of a more general consequence acceptable?", Mark Thomas, "Breaking the Chain of Causation: Reasonable Foreseeability and the 'Exact Form' of a Subsequent Act: R v A [2020] EWCA Crim 407; [2020] 1 WLR 2320" (2020) 84 *Journal of Criminal La*w, p. 626. See also *Wayne LaFave*, *supra* note 28, pp. 263–264;

[39] Johannes Himmelreich, "Responsibility for Killer Robots" (2019) 22 *Ethical Theory and Moral Practice*, pp. 743–744.

There are several reasons why modern AI exacerbates this lack of foresight. Unpredictability is a major one,[40] but one which is also unavoidable for the tasks we expect our AI to accomplish. AI systems such as those installed in AVs are, by definition, faced with "a world filled with uncertainty, volatility, and flux",[41] a dynamic setting that must be navigated by the AI independently and flexibly. It is not a task which can be hard-coded by programmers: indeed, this is the reason ML techniques are used in the first place.[42] However, lower predictability is an unavoidable consequence of such techniques that we must grapple with. This is much in contrast to more traditional rule-based AI, which has "one major virtue: it is always clear why the machine makes the choice that it does, because its designers set the rules".[43] While we can expect a user or designer to be able to foresee the behaviour of rule-based AI for the purposes of *mens rea*, this is much more challenging for modern AI based on ML.

Lack of predictability is exacerbated in situations where the machine is allowed to actively learn in the field. On the one hand, this is beneficial since it allows the machine to improve its performance over time.[44] However, it is evident that, by giving the system the opportunity to continue developing after being released as a product, predictability further decreases, with evident repercussions on the allocation of responsibility.[45] In his initial publication, Matthias provided many prototypical illustrations to this effect. One example features a pet robot which "learns" to gallop to reduce battery consumption and ends up ramming violently into a child. Matthias comments how one could view this incident as "an *unforeseeable*

[40] Ugo Pagallo, "When Morals Ain't Enough: Robots, Ethics, and the Rules of the Law" (2017) 27 *Minds and Machines*, p. 634.

[41] David Leslie, *Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector* (Alan Turing Institute 2019), p. 30.

[42] S.J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach* (3rd edn, Pearson 2010), p. 693.

[43] Boer Deng, "The Robot's Dilemma: Working out How to Build Ethical Robots Is One of the Thorniest Challenges in Artificial Intelligence" (2015) 523 *Nature*, p. 25.

[44] See eg D. Sculley and others, "Machine Learning: The High Interest Credit Card of Technical Debt", *SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)* (2014), p. 3.

[45] European Commission, "Report on the Safety and Liability Implications of Artificial Intelligence, the Internet of Things and Robotics" (2020) COM(2020) 64 final, p. 9.

development, which occurred due to the adaptive capabilities of the robot, so that nobody can be justly said to be responsible".[46]

Complexity and opacity are another aspect unique to modern AI which obfuscates predictability. Both refer to the situation where it is simply not knowable – *even for its creators*, and thereby much less so for its lay-users – how an AI system functions and makes decisions. An AI system's architecture may be so complex, featuring multiple interacting subsystems, that understanding how the overall product functions may be impossible.[47] Wallach & Allen submit that expecting "operators to anticipate the actions of intelligent systems becomes more and more unreasonable as the systems and the environments in which they operate become more complex".[48] This complexity often comes in combination with opacity, a notable characteristic of many ML systems, particularly deep neural nets. These AI systems are often referred to as black boxes, described by the European Commission as systems "that do not allow cognitive access to how they have arrived at a particular output, or what input factors or a combination of input factors have contributed to the decision-making process or outcome".[49] Black box systems are intrinsically intractable, even for experts and their own designers,[50] and are commonly used in AVs.[51] An argument could therefore potentially be made before a court that it was impossible (in a non-hyperbolic sense) for the accused to foresee that the AV would commit the offence.

Fortunately, there are methods being developed in the AI domain to reduce this intractability. For instance, eXplainable AI (XAI) specifically researches methods to render modern AI more transpar-

[46] Matthias, *supra* note 3, p. 176 (emphasis in original).

[47] King, *supra* note 20, p. 95.

[48] Wendell Wallach and Colin Allen, "Framing Robot Arms Control" (2013) 15 *Ethics and Information Technology*, p. 132.

[49] Horizon 2020 Commission Expert Group to advise on specific ethical issues raised by driverless mobility (E03659), *Ethics of Connected and Automated Vehicles: Recommendations on Road Safety, Privacy, Fairness, Explainability and Responsibility* (Publication Office of the European Union 2020), p. 49.

[50] Vijay Arya and others, "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques" (2019) <http://arxiv.org/abs/1909.03012>, p. 1.

[51] Will Knight, "The Dark Secret at the Heart of AI" (*Technology Review*, 2017) <www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai> accessed 2 July 2020.

ent and explainable.[52] Such efforts are being pushed for many AI applications, including AVs,[53] and would somewhat mitigate the abovementioned obstacle with regards to the accused's cognitional element. Nevertheless, even with XAI, understanding the system may still require some study or training. Particularly lay-users (ie, likely the large majority of persons purchasing an AV) will have no background in the technology and no desire to invest time and effort to allow such understanding – they will simply want their car to drive them to their destinations. This potentially allows deniability, ie, cases where the user *could* have known the functioning of the AI system but, in practice, *did* not know (or so they might claim). Even worse, persons may be *incentivised* to learn as little as possible of their AV if this reduces their risk of criminal liability. As was observed by Williams before the British House of Lords, the current situation "provides a great incentive for human agents to avoid finding out what precisely the ML system is doing, since the less the human agents know, the more they will be able to deny liability for both these reasons".[54]

2.3 *The Issue of Generic Risk*

One may argue at this point that awareness of risk need not necessarily be tied to a thorough understanding of the AI mechanics at play. The owner of a car with defective brakes of the previous example, for instance, does not need to have studied theoretical hydraulics to understand that driving that car would very likely result in harm to others – sufficiently so for *mens rea* to be established. Two aspects, however, complicate this position slightly with respect to AI.

First, as discussed above, the dynamicity and unpredictability of modern AI outputs makes it less clear whether the accused could foresee *particular* result, or if they were merely aware of a generic and vague possibility that something might go wrong. It is submitted that in a large number of cases, the accused's awareness will be limited to

[52] Amina Adadi and Mohammed Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)" (2018) 6 *IEEE Access*, p. 52138.

[53] House of Lords, "Select Committee on Artificial Intelligence, Report of Session 2017–2019, AI in the UK: Ready, Willing, and Able?" (2018) HL Paper 100, 16 April 2018, p. 37.

[54] Rebecca Williams, "Lords Select Committee, Artificial Intelligence Committee, Written Evidence (AIC0206)" (2017) <http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/70496.html#_ftn13> accessed 7 May 2022.

the latter. Perhaps a developer was aware of some unidentified edge cases where the AI might malfunction but chose to release the product anyway for expediency; or perhaps users might be aware that their AV's performance is not as high in the rain but callously activate the program anyway. In both situations agents are aware of *a* risk, but not any risk in particular. This may be problematic depending on the legal system's conception of risk and on the level of specificity required: does the accused need to be able to predict a *specific* consequence (''this family might get killed''), a *category* of harm (''I might hit a pedestrian''), or simply a risk in general (''something might go wrong'')?[55] The consequences of driving around with defective brakes are envisageable and restricted; what a failing AI could do is theoretically limitless.

Second, and related to the previous point, at what point does a general sense of risk become blameworthy? All users of tools are aware of the chance of something going wrong, as all machines have failure rates – no machine is infallible. Accepting this probability however does necessarily not amount to *mens rea* – otherwise, *any* mechanical failure would trigger what would basically be a strict liability regime. Often, a guardrail is placed in the form of the requirement that the risk must be unreasonable and/or substantially likely to occur,[56] although the details differ per jurisdiction.[57] As such, depending on how *culpa* is formulated, awareness of some indeterminate risk of AI failure (the statistical probability of which might also not be known) could be insufficient for establishing *mens rea*. In fact, if an argument could be made that no reasonable person *could* have known owing to the AI's complexity and opacity, even negligence charges would be barred.[58]

---

[55] See Neha Jain, ''Autonomous Weapons Systems: New Frameworks for Individual Responsibility'' *in* Nehal Bhuta and others (eds), *Autonomous Weapons Systems: Law, Ethics, Policy* (Cambridge University Press 2016), p.317.

[56] Jeroen Blomsma and David Roef, ''Forms and Aspects of Mens Rea'' *in* J. Keiler and D. Roef (eds), *Comparative concepts of criminal law* (Intersentia 2015), pp. 186–192. See eg Model Penal Code, §2.02.(2).(c), which requires the person to have disregarded a ''*substantial and unjustifiable risk''*.

[57] J.D. Ohlin, ''Targeting and the Concept of Intent'' (2013) 35 *Michigan Journal of International Law*, 103. See eg Blomsma and Roef, *supra* note 56, pp. 183–185, comparing *dolus eventualis* in the Netherlands and Germany, the former requiring a ''considerable chance'' of the risk materialising, while the latter only requiring that this chance exists.

[58] Hallevy, *supra* note 4, p. 205.

2.4 *A Lack of Control*

The second major strain of objections against assigning responsibility for acts of AI agents relates to the *control condition*.[59] Control is not an explicit legal element a prosecutor must prove, but many view as foundational to one of the fundamental purposes of criminal law: punishing only *culpable* conduct. Vincent, expanding upon Hart's 1968 seminal work on different types of responsibility,[60] explains how one can only be responsible for an outcome if one had *causal control* over its occurrence.[61] This sentiment was reflected in Matthias' article, who remarked that a responsibility gap occurs when "nobody has enough control over the machine's actions" to make a justifiable attribution of responsibility, since control is a "necessary condition" for it.[62] This requirement is broadly echoed in literature[63] and also the main reason why, in the analogue discussion concerning the responsibility for war crimes committed by AI, commentators gravitate toward imposing a requirement of direct control over the AI's decisions, hoping that thereby, the control condition can be fulfilled.[64]

One relatively expedient solution that seemingly addresses this problem – without completely negating the benefits of having AI in the first place by simply forbidding autonomous decision-making entirely[65] – is to impose a duty to intervene on a specific actor, eg, an operator. This operator would then act as supervisor for the system, who can take over for the machine in case the latter malfunctions or encounters difficulties. This way, at least in theory, one would obtain

---

[59] Himmelreich, *supra* note 39, p. 735.

[60] cf H.L.A. Hart, *Punishment and Responsibility: Essays in the Philosophy of Law* (Clarendon Press 1968), pp. 211 ff.

[61] Nicole Vincent, "A Structured Taxonomy of Responsibility Concepts" *in* Nicole A Vincent, Ibo van de Poel and Jeroen van den Hoven (eds), *Moral Responsibility: Beyond Free Will and Determinism, vol 27* (Springer Netherlands 2011).

[62] Matthias, *supra* note 3, pp. 175–182.

[63] Eg Himmelreich, *supra* note 39, 732–736; Robin Geiß and Henning Lahmann, "Autonomous Weapons Systems: A Paradigm Shift for the Law of Armed Conflict?" in Jens David Ohlin (ed), *Research Handbook on Remote Warfare* (Edward Elgar 2017), p. 393.

[64] Jonathan Kwik, "A Practicable Operationalisation of Meaningful Human Control" (2022) 11 *Laws*, 15–16.

[65] *See eg* A.L. Schuller, "Artificial Intelligence Effecting Human Decisions to Kill: The Challenge of Linking Numerically Quantifiable Goals to IHL Compliance" (2019) 15 *Journal of Law and Policy for the Information Society*, p. 115.

the benefits of autonomous decision-making while also maintaining the human as a risk management tool.[66] For AVs, the most obvious candidate for this role is the driver already sitting behind the wheel.

However, this solution encounters significant problems in practice. As our control condition requires, responsibility can only be imputed upon such an intervening operator if this person has actual *meaningful* control over subsequent events. Depending on the situation, this may not be the case. There are several reasons for this. First, it has been scientifically established that humans are very ineffective supervisors, and passive monitoring usually lulls persons into reduced states of attentiveness and situational awareness.[67] This effect has been observed many times with respect to autopilots. Perrow recounts that often, "when the pilot is suddenly and unexpectedly brought into the control loop (in other words, participates in decision making) as a result of (inevitable) equipment failure, he is disoriented … The sudden appearance of several alarms, all there for safety reasons, leads to disorientation".[68] Even for several seconds after this person is expected to have taken back control, they may not possess the capacity to properly and reasonably act to avoid an imminent harm, and thereby not truly be in control.

Related to this is a lack of time. If an operator is expected to take over control from an AI to prevent harm, they must be provided the required opportunity to do so. Depending on how imminent a disaster is, such as a collision with a pedestrian, it might simply be a superhuman ask to demand them to respond in time. If the decision must be made in mere seconds or even milliseconds, "actual control over the system's actions may be no more than an illusion".[69] These two factors in combination make us question whether a human supervisor could truly have *de facto* control over a vehicle, even when

[66] A.J. Lohn, "Estimating the Brittleness of AI: Safety Integrity Levels and the Need for Testing Out-Of-Distribution Performance" (2020) <http://arxiv.org/abs/2009.00802>, p. 6.

[67] R. Parasuraman and others, "A Model for Types and Levels of Human Interaction with Automation" (2000) 30 *IEEE Transactions on Systems, Man, and Cybernetics*, p. 291.

[68] Charles Perrow, *Normal Accidents: Living With High-Risk Technologies* (Basic Books 1984), p. 132.

[69] Geiß and Lahmann, *supra* note 64, 378. In addition, the required time may be idiosyncratic. A person of age, for instance, will require a longer period to act with the same level of effectiveness as younger drivers.

such control has technically been ceded back by the AV. Imputing responsibility for what results, then, is problematic.[70]

2.5 *The Problem of Distance and Many Hands*

In addition to the actual end-users (such as drivers), much discussion has also surfaced with respect to criminal liability for actors earlier in the production chain, such as programmers, designers, or the sellers and distributors (let us call these *prior chain actors*, ''PCA''). For the purposes of the current article, we will focus on PCAs' criminal liability over alternative forms of liability, such as tort liability. We have already seen above that establishing *mens rea* would be possible in the presence of intent to create or distribute a product meant to deliver a harmful event.[71] More problematically for PCAs, their relative temporal and physical distance from the event raises issues of establishing causality. Gogarty & Hagger remark that even negligence-based regimes might be limited by ''salient considerations of causal, physical and circumstantial proximity which seek to place a reasonable constraint on unfair or burdensome duties being imposed on those who are simply too far removed from the act that caused harm''.[72]

Additionally, unlike the singular driver, PCAs often form part of large corporations and organisations with interconnected departments and hierarchies. Pinpointing and proving the cause of a specific failure will be very challenging, exacerbated by the fact that (groups of) individuals in this organisation can easily shift blame to other persons or departments for the failure.[73] The failure may also not be ''caused'' by a single mistake, but manifested only as a result of unforeseen interactions between several of them, further complicating the process of focalising blame.[74] This is often referred to as the *problem of many hands*. Coined by Thompson in 1980, the term is used to refer to the dilution of intent, knowledge and decision-making

[70] See Vincent, *supra* note 61, p. 21.

[71] Subsection 2.1.

[72] B. Gogarty and M.C. Hagger, ''The Laws of Man Over Vehicles Unmanned: The Legal Response to Robotic Revolution on Sea, Land and Air'' (2008) 19 *Journal of Law, Information and Science*, p. 123.

[73] Williams, *supra* note 54.

[74] J.M. Beard, ''Autonomous Weapons and Human Responsibilities'' (2014) 45 *Georgetown Journal of International* Law, p. 651.

power over a network of actors and groups.[75] It is an issue criminal law has struggled with in general and one of the primary reasons why regimes such as responsibility for legal persons and corporations were invented.[76] In 1996, Nissenbaum developed this theory further and found it to be particularly salient in software development processes. As she explains, software "are the products not of single programmers working in isolation but of groups or organisations, typically corporations … which frequently bring together teams of individuals with a diverse range of skills and varying degrees of expertise … Consequently, when a system malfunctions and gives rise to harm, the task of assigning responsibility … is exacerbated and obscured".[77]

Thus, in a many hands scenario, it "may not be obvious who is to blame because frequently its most salient and immediate causal antecedents don't converge with its locus of decision making".[78] For example, an AV crash might have been "caused" by a combination of some mischievousness by data labellers, a reckless oversight by a programmer, laziness of quality control staff, a mechanical defect with the AV's forward sensor, and a desire for quick profit by the managing board.[79] This raises problems not only with demonstrating causality, but also the accused's cognition: did they foresee their seemingly insignificant act as likely to cause a deadly accident, perhaps half a year from then? Even negligence claims might be difficult to pursue in this light if the defence can demonstrate that no reasonable person could have foreseen that result.

III DRAWING BRIGHT LINES: THE SINGAPOREAN, FRENCH, AND BRITISH APPROACHES

In the previous section, we have examined a range of troubles that make allocating liability for criminal offences performed by artificial agents challenging – at least, if one wishes to do so fairly. One option would be to just close our eyes to the discussion in Section II and insist that AI systems are nothing new under the sun. One could, for

[75] See Dennis Thompson, "Moral Responsibility and Public Officials: The Problem of Many Hands" (1980) 74 *American Political Science Review*.

[76] Williams, *supra* note 54.

[77] Helen Nissenbaum, "Accountability in a Computerized Society" (1996) 2 *Science and Engineering Ethics*, 28–29.

[78] *ibid* p. 29.

[79] See eg *ibid* p. 30.

instance, simply impose a duty to retake control and make the driver responsible for anything that occurs after this moment (disregarding the lack of *de facto* control), or insist that the accused should have known of the risks (when even its creators might not fully understand the AI's functioning). Such approaches would however fundamentally be in opposition to the basic philosophy of criminal justice and our intuition on fairness. "The idea of punishing only those with a guilty mind is well grounded in natural justice and human rights … the fact that 'no man ought to be punished, except for his own fault' is a clear maxim of natural justice." [80] The aim of new legal regimes as those which we will discuss in this section, then, should be to allow a fair administration of criminal justice, whilst addressing the issues we have identified in Section II.

The following subsections will focus on three countries: Singapore, the UK, and France, ie, the first three countries to have enacted (or proposed) hard-law regulation on criminal liability for AI misbehaviour. First, we will analyze two Singaporean proposals to amend the Singaporean Penal Code: the Singapore Penal Code Review Committee Report of 2018 and the recommendations on Criminal Liability, Robotics and AI Systems of the Singapore Academy of Law's Law Reform Committee published in February 2021. Then, we will briefly outline the French *ordonnance* of April 2021, which amended the French Road Code[81] by specifically adding a chapter on criminal liability applicable to the use of a vehicle with delegation of driving functions. Next, Subsection 3.3 will focus on the Joint Report on Automated Vehicles drafted by the Law Commission of England and Wales and the Scottish Law Commission ("the UK Law Commissions"), which was released at the end of January 2022.[82] When relevant, we will compare the Joint Report to the aforementioned Singaporean and French proposals. As it will be shown, one recurrent feature of the proposals that are analysed in this section is the act of

[80] Thompson Chengeta, "Accountability Gap: Autonomous Weapon Systems and Modes of Responsibility in International Law" (2016) 45 *Denver Journal of International Law & Policy*, p. 19.

[81] Ordinance of 14 April 2021 n. 2021-443 ("Responsabilité pénale applicable en cas de circulation d'un véhicule à délégation de conduite (articles L.123-1 à L.123-4")) which added a new chapter in the Title 2 of the French Road Act ("Code de la route").

[82] Law Commission of England and Wales Report (Law Com No 404, 2022), Scottish Law Commission (Law Com No 258), "Automated Vehicles: Joint Report" [4.1.] (hereinafter "LCR").

"drawing lines". On one side of the line, we find liability, and on the other, immunity.

3.1 *Singapore: the (Criminal) Rule of Law Hub*

Seemingly, Singapore is seeking to establish itself as an AI "rule of law hub",[83] by means of introducing regulation "to attract and encourage AI innovation".[84] Indeed, the proposals analysed in this subsection are a paramount example of the Singaporean strive to become key normative players in the field AI. Already back in 2018, the Singapore Penal Code Review Committee ("PCRC")[85] acknowledged that "[b]eing the global first-mover"[86] might "impair Singapore's ability to attract top industry players in the field of AI".[87] Nevertheless, the PCRC advised the Singaporean government to "actively explore and develop a suitable framework to address the issue of criminal liability for harm caused by computer programs … in the broader context of Singapore's developing regulatory framework for AI".[88-89] In this subsection, we will examine two different propositions. First, the PCRC Report of 2018, specifically the proposal to introduce two new offences relating to computer programs. Second, the Singapore Academy of Law's ("SAL") Law Commission Report on Criminal Liability, Robotics and AI Systems of 2021. As will be shown, in contrast to the UK and French examples, which are focused on autonomous driving,[90] both Singaporean initiatives have a wider scope of application. That is, they discuss criminal liability for *any* harmful act involving an AI system, and not just in the field of autonomous driving. Hence, they represent the first attempts at building a general framework of criminal liability for to AI crime.

---

[83] Simon Chesterman, *We, the Robots? Regulating Artificial Intelligence and the Limits of the Law* (CUP, 2021), p. 5.

[84] *ibid*

[85] Singapore Penal Code Review Committee, *Report*, August 2018 (hereinafter "*PCRC*").

[86] *ibid* p. 29.

[87] *ibid*

[88] *ibid*

[89] The PCRC Report includes AI in the term "computer programs". See *ibid* p. 27.

[90] See Subsections 3.2. and 3.3.

3.1.1 *The Singapore Penal Code Review Committee's Report*

The PCRC was established by the Singaporean Ministry of Home Affairs and Ministry of Law in 2016 to review the Singapore Penal Code and make recommendations on how to reform it. It completed its review in 2018 and released a comprehensive report where, amongst other things, it suggested the introduction of two new offences which regulate the attribution of criminal liability in cases of harm caused by computer programs. For the sake of clarity, we will refer to them as Offence A and Offence B. Why the analysis of the PCRC Report is relevant is twofold. Firstly, it contains the first (and only) draft formulations of negligence offences specifically tailored to AI systems. Secondly, the SAL Report, which will be analysed further, builds upon the findings of the PCRC Report.

Let us now move to the analysis of the contents of Offence A and Offence B. Offence A would be structured as follows:

(1) Whoever makes, alters or uses a computer program shall be punished with imprisonment for a term which may extend to one year, or with fine which may extend to $5,000, or with both.
(2) For the purposes of this section, a person uses a computer program if he causes a computer holding the computer program to perform any function that —
    (a) causes the computer program to be executed; or
    (b) is itself a function of the computer program.
(3) For the purposes of this section, a computer program is under a person's care if he has the lawful authority to use it, cease or prevent its use, or direct the manner in which it is used or the purpose for which it is used.[91]

This offence would impose liability on two categories of subjects: first, those who make and alter computer programs (programmers); second, on those who use them (operators).[92] Specifically, it would address conducts of "risk-creation"[93] regardless of the verification of harm. In other words, it would constitute an instance of a crime of endangerment.[94] In case harm were to manifest as a consequence of said risk, whether resulting in physical injury or death, the application of

[91] PCRC, *supra* note 85, p. 30 (emphasis added).

[92] *ibid*

[93] *ibid*

[94] Crimes of endangerment punish acts or omissions that create significant risk that someone will suffer harm, *regardless of whether the risk is actualized*. They consist of a "failure of proper concern". See Anthony Duff and Tatiana Hörnle , "Crimes of Endangerment" in Kai Ambos and others (eds), *Core Concepts in Criminal Law and Criminal Justice* Volume 2 (CUP: 2022); Anthony Duff, "Criminalizing Endangerment" (2005), 65 *La. L. Rev*., p. 944.

other offences of the Singapore Penal Code would be triggered (such as articles 304A or 337).

If, on one hand, *actus reus* elements of endangerment offences do not seem to pose particular issues at first glance, on the other, it is debated whether the *mens rea* connection is one of strict liability or of fault. As a matter of fact, in a more culpability-principle-compliant perspective, the offender shall be liable for being *indifferent* to the risk they created, that is, they shall display an attitude of not caring for the legally protected interests.[95] This concern is addressed in the proposed wording of the PCRC, which states clearly that the offender shall act ''rashly or negligently or knowingly''. The PCRC here draws a line: the user should have *known* that there was a risk of harm for one's life or physical integrity. This awareness is referred to as ''rashness''. However, as noticed by the PCRC, Offence A would leave out scenarios where (1) the peril impacted legal goods other than human life and human integrity, and (2) the ''user'' was not aware that the (specific) harm will occur, ''either because the program is capable of learning new behaviours on its own or because the program is designed to act random''.[96] In other words, the PCRC argue that a *lacuna*, consisting of (1) + (2), would arise.

Let us try now to situate the Singaporean concept of ''rashness'' in the discussion on the concept of negligence conducted at Subsection 2.1. Rashness is a form of culpability regulated in the Singapore criminal legal system and which, as negligence, arises from a ''failure to exercise a degree of care and caution expected of the actor''.[97] Though, rash acts usually are ''of a more active and exceptional nature, where the actor acts imprudently or impetuously without taking the required steps to ensure that the act is carried out safely''.[98] Negligence, instead, typically arises ''from routine acts which, though unexceptional in and of themselves, are nevertheless commonly understood to give rise to some degree of danger''.[99] In other words, it is a matter of expectations: if drivers cause an accident because of a failure to pay attention to the surroundings, they acted

[95] *ibid*

[96] PCRC, *supra* note 85, p. 30.

[97] Sundram Peter Soosay, ''The Work of Many Hands: the Continuing Confusion over Section 304A of the Singapore Penal Code'' (2015), *Singapore Journal of Legal Studies*, p. 144.

[98] *ibid* p. 144.

[99] *ibid*

negligently; if drivers instead cause an accident because of speeding up at an intersection at the prospect of a red light, they acted rashly. According to some, the difference between rashness and negligence lies in the consciousness of risk. When this is present, then the offender is acting rashly. When it is absent, the offender is acting negligently.[100]

What solution does the PCRC propose to address this apparent *lacuna*? They suggest the introduction of Offence B:

> (1) Where a computer program —
>
> (a) produces any output, or
> (b) performs any function,
>
> that is *likely* to cause any hurt or injury to any other person, or *any danger* or annoyance to the public, and the computer program is *under a person's care*,
>
> if that person *knowingly* omits to take *reasonable steps to prevent* such hurt, injury, danger or annoyance, he shall be punished with imprisonment for a term which may extend to one year, or with fine which may extend to $5,000, or with both.[101]

Offence B seems to cover cases where the person overseeing the computer program was *not* aware of the existence of *any* kind of risk, a situation which we referred to above as scenario (2). Indeed, the proposed formulation of Offence B does not tie the duty of care, consisting in taking reasonable steps to prevent harm, to the *knowledge* of the risk that the computer program is *likely* to cause any hurt or injury to any other person, or any danger or annoyance to the public. Knowledge is attached only to the conduct of not acting upon the risk of danger with preventive measures. In other words, differently from Offence A, Offence B seems to entail that a user could be liable even if the risk of harm was an objective and intrinsic characteristic of the computer program, ie, one that is *independent from any subjective evaluation of the culpable agent*, and he or she did not act upon this characteristic to mitigate the risk.

Now, one could ask herself what the scope of application of the word "likely" is. Does it entail the knowledge that the computer has a 51% probability to cause harm? Or is the threshold higher? More-

[100] *ibid* p. 145.

[101] PCRC, *supra* note 85, pp. 31–32.

over, are all learning AI inherently "likely" to cause harm? If not, what characteristic of a learning AI would make it "likely" to hurt/injure/etc.?

Lastly, Offence B supposedly also addresses lacuna (2), as it expands the scope of application to "*any* danger or annoyance to the public".[102] Thus, this represents an extremely broad formulation. Even if one were to interpret Offence B as demanding that the agent possessed the knowledge of the likeliness that the AI system would cause any danger or threat to the public, it would be extremely hard, if not impossible, for any reasonable agent to fulfil such a high threshold of knowledge as one including the threat of any danger or annoyance to a public. In conclusion, it appears that with Offence B the drafters of the Report overstepped that *mens rea* line that they tried to draw with Offence A.

3.1.2 *The Singapore Academy of Law's Report*

Let us move now to the second proposal: the Report on Criminal Liability, Robotics and AI Systems of the SAL's Law Reform Committee ("LRC"), which was published in February 2021.[103] The SAL is a private body – differently from the PCRC – established in 1988 with the purpose of making Singapore the "legal hub of Asia".[104]

The SAL Report examines potential risks posed to humans and property by the use of autonomous robotic and AI systems ("RAI"). It focuses on situations in which harm arises and on whether, and how, Singaporean criminal laws should apply, and criminal liability attributed.[105] Notably, the drafters of the report acknowledge the variety of potential RAI applications (each entailing differing sources and levels of risk, responsibility, and benefits) which makes a "one size fits all" to criminal liability unpracticable.[106] For these reasons, the Report's analysis is not sector-based but is instead conducted taking two factors into account: first, whether or not there was a human "involved in operating, affecting, or overseeing the RAI

[102] *ibid* (emphasis added).

[103] Singapore Academy of Law Reform Committee, "Report on Criminal Liability, Robotics and AI Systems" (2021), DC 345.595704–dc23 (hereinafter "SAL").

[104] The SAL is led by a Senate, headed by the Chief Justice and comprising of the Attorney-General and the Supreme Court Bench; and its members include over 14,000 legal professionals or academics. See <www.sal.org.sg>.

[105] SAL, *supra* note 103, [1.3].

[106] *ibid* [1.4].

system''; second, ''where such a human is involved, whether they intended or knew the harm would occur''.[107]

3.1.2.1 *New Legal Actors, Take One: the Singaporean ''User-In-Charge''.* As mentioned above, the LRC argues that whether – and on whom – criminal liability should be imposed is likely to be a function of: the severity and risk of actual or potential harm inherent in the use of the system in the relevant context; the level of automation of that system; and the degree of human oversight over, and involvement in, the system's decision-making (if any).[108] Focusing on the last two, according to the LRC the first issue would be to identify the ''user-in-charge''.[109] In cases of ''partial automation'', ie, where the level of automation is lower than the level of human oversight exercised, the user-in-charge would be the subject who ''directly controls or is responsible for determining the actions of the RAI systems''. In cases of highly (yet not fully) automated RAI systems, the user-in-charge would either be the subject who bears ultimate responsibility for deciding on or approving a particular action, the one who retains oversight over the system's decision-making process, or the one who is under a specific duty to intervene to control the system's action in a given scenario.[110]

At this point, we must underline how the term ''user-in-charge'' adopted in the SAL Report is the same as the one put forth by the UK Law Commissions in their Joint Report, which will be examined in Subsection 3.3. Notably, the LRC explicitly addresses this overlap and states that ''[w]hile utilising the same term, the definition of 'user-in-charge' adopted here differs from that utilised by the UK Commissions in the specific context of automated vehicles (although its 'users-in-charge' would equally fall within the definition utilised here)''.[111] In other words, all ''UK'' users-in-charge would qualify as ''Singaporean'' users-in-charge, but not the other way around.

Focusing now on non-intentional harms, the SAL Report mentions that according to the Singapore Penal Code negligence is established when to conditions are fulfilled: (a) determining what an objective ''reasonable person'' would do in a given circumstance, and

[107] *ibid* [1.5].

[108] *ibid* [4.1].

[109] *ibid*

[110] *ibid* [4.3].

[111] *ibid* n. 34., p. 26.

(b) proving that the standard was breached in the specific case.[112] When it comes to harm caused by a RAI, which falls within the scope of already existing negligence-based offences, it would be up to the courts to "apply or adapt existing criminal negligence standards, or – in the absent of precedent – define new one".[113] Moreover, the reasonable conduct standard could be set by new legislation, through the creation of a new negligence-based offence to cover all negligent conducts which lead to harms from RAI systems. Thus, the risk of such a general applicable provision is that it could prove insufficient to capture conducts of RAI systems which have never happened before, ie, for which "existing precedents are inappropriate or for which there is no existing precedent at all".[114]

What is more, the LRC reflects on introducing technology or sector-specific standards of conduct through legislation. One of the examples mentioned in the SAL Report is the one of AVs: the LRC suggests that legislation might provide for certain circumstances in which the user-in-charge must take control of the vehicle, such as when a road is closed temporarily due to a traffic accident.[115] In this sense, the approach of the SAL Report differs from the one undertaken by the UK Law Commissions. As we will see, the latter, rather than focusing on *external circumstances* (eg, an accident) and their impact on the duty of the operator to intervene, focuses on the fitness of the single Autonomous Driving System feature of the vehicle, to be certified through an authorization scheme. Indeed, according to the UK approach, the user-in-charge will not be liable for *any* "dynamic driving offence"[116] or civil penalty committed when such feature is engaged.

3.1.2.2 *Failures and Gaps*. As mentioned above,[117] specific features of modern AI could lead to situations where harm is caused, yet no negligent conduct by the user-in-charge can be identified, ie, negligence failures. It is relevant to note here how the LRC attentions aspects which are usually disregarded by the scholarly debate. Specifically, it highlights the importance of every stage of the AI

[112] Singapore Penal Code 1871, S 26F(1) PC.

[113] SAL, *supra* note 103, [4.26].

[114] *ibid* [4.27].

[115] *ibid*

[116] See Subsection 3.2. for a definition of dynamic driving offences.

[117] See Section II.

deployment process (data preparation, training of the mode, choosing the relevant model[s], the environment where the RAI system is deployed) as probable causes of the realisation of (criminally relevant) harm.[118] Moreover, the LRC points out that harm might be caused not only by the architecture (ie, the code) of the RAI system, but also by the quantity, quality, and accuracy of the training data. One should also take into account, on the one hand, the relevance of comparing the environment in which the system was trained with the one in which it was deployed and, on the other, what real-world data was collected by the RAI system at the time the harm was committed.[119]

The Committee identifies three causes which could lead to negligence failures. First, the fact that multiple players are involved in the AI deployment process, ie, the many hands problem. Note however that, as highlighted above,[120] the many hands problem can manifest both as a phenomenon that causes an *intrinsic* problem with *mens rea* (where knowledge simply does not exist in any PCA because the risk was completely unforeseeable) and a more *evidentiary* problem (where prosecutors struggle to focalise liability due to the magnitude of PCAs involved). Second, the different types of RAI systems; and third, the ability of a RAI to learn from its surroundings and produce unexpected and unexplained harmful outcomes. The last two causes clearly refer to the epistemic problems discussed *supra*,[121] where the knowledge component of *mens rea* is absent because of the system's complexity, opacity, and ability of "online learning".

How shall these failures be addressed? According to the LRC, criminal negligence might not (always) be the answer. They suggest four alternative criminal liability mechanisms, which we will briefly analyse here. The first one is the creation of a new form of legal personality (or "personhood") for RAI systems, such that criminal liability could be imposed on the RAI system itself. The LRC eventually discards this option, since it considers the arguments against separate personality for RAI systems more compelling.[122] The second and the third alternatives, instead, are the offences which were the-

[118] SAL, *supra* note 103, [4.32].

[119] *ibid*

[120] See above Subsection 2.5.

[121] See above Subsection 2.2.

[122] SAL, *supra* note 103, 4.47, 38.

orized in the PCRC Report (ie, Offence A and Offence B).[123] Here the LRC notices that even though these offences could indeed address negligence failures, they still do not contour with enough precision the perimeters of the duty of care over the "computer program". In other words, more effort is needed in specifying exactly what constitutes "a rash or negligence act or failure to take reasonable steps in any given circumstance".[124] The fourth alternative is to use the workplace safety legislation of Singapore as a model. Said legislation imposes on employers a duty "to take, so far as it is reasonably practicable, such measures as are necessary to avoid harm"[125] in the workplace. This duty could be imposed on the subject who displays most "proximity" the RAI system, taking also into consideration its resources to take action and to change future outcomes. This subject, as it will be explained below, is akin to the Automated Driving System Entity as envisioned by the UK Law Commission.[126]

3.2 *France: the Ordonnance "Responsabilité pénale applicable en cas de circulation d'un véhicule à délégation de conduite"*

In 2021, the French Parliament adopted an *ordonnance* which amended the French Road Act to address criminal liability for traffic offences committed by AVs. Specifically, it added a new article 123-1 to the Road Code, which reads as follows:

> The provisions of the first paragraph of article L. 121-1 are not applicable to the driver for violations resulting from the operation of a vehicle whose driving functions are delegated to an automated driving system, when this system exercises, at the time of the violation and under the conditions provided for in I of article L. 319-3, the dynamic control of the vehicle.[127]

This article, then, excludes the application of article L. 121-1 of the French Road Code, which provides that "[t]he driver of a vehicle shall be criminally liable for violations committed while operating said vehicle",[128] to the "driver" present on an AV, provided that the

[123] See above Subsection 3.1.1.

[124] SAL, *supra* note 103, [4.54].

[125] *ibid* [4.58].

[126] See Subsection 3.3.

[127] French Road Act ("*Code de la route*"), art L. 123-1 (author's translation).

[128] *ibid*

driving functions and the dynamic control of the vehicle had been correctly delegated to the AI.

As we mentioned above, we can identify a recurrent element, ie, the act of ''drawing a bright line''. With regards to France, the ''bright line'' is drawn by establishing that, in order for the immunity clause to operate, the ADS must have had *dynamic* control of the driving functions when the offences were committed.[129] Yet the *ordonnance* itself does not give a definition of dynamic control,[130] which can be found instead in a decree adopted on June 29, 2021, at article 2. Dynamic control is defined as ''[t]he performance of all real-time operational and tactical functions required to move the vehicle'', which include ''control of the lateral and longitudinal movement of the vehicle, monitoring of the road environment, response to events in road traffic, and preparation and reporting of maneuvers''.[131]

According to L.123-1, par.2, the drivers, on their part, must always be in a position to respond to a request to take control by the automated driving system (''ADS'').[132] This provision severely cripples the scope of application of the immunity clause in par.1. As noted above,[133] *reprise* clauses are very delicate since improper formulation bears the risk of making the immunity clause functionally void (by demanding superhuman feats from users). Moreover, similarly to the British approach, according to L.123-1, par.3., accepting the *demand de reprise,* or failing to do so, provided that the transition period has passed, will lead to the re-expansion of the scope of application of article L.121-1 of the Code, ie, to criminal liability.

According to some, this newly introduced immunity clause works as a mere *reconnaissance* of a conclusion which could have been reached by applying standard principles of criminal law, specifically the rules on negligence.[134] The real turning point, instead, would be article L319-3, which regulates the conditions for the correct *activation* of the dynamic control of the vehicle by the AI system. Indeed, article L319-3 provides that

[129] See Marta Giuca, ''Disciplinare l'intelligenza artificiale. La riforma francese sulla responsabilità penale da uso di auto a guida autonoma'' (2022), 2 *Archivio Penale*.

[130] *ibid* 22.

[131] Decree of 29 June 2021 n. 2021-873, TRAT2034544D.

[132] *Code de la route*, article L. 123-1, II.

[133] See above Subsection 2.4.

[134] Giuca, *supra* note 129, p. 23.

I.- The decision to activate an automated driving system is taken by the driver, who has been *previously informed* by the system that it is capable of exercising dynamic control of the vehicle in accordance with its conditions of use.
II.- When its operating state no longer allows it to exercise dynamic control of the vehicle or when its conditions of use are no longer met or when it anticipates that its conditions of use will probably no longer be met during the execution of the maneuver, the automated driving system must:

(1) *Alert* the driver;
(2) *Make a request* to regain control;
(3) initiate and execute a minimum risk maneuver if control is not regained at the end of the transition period or in the event of a serious malfunction.(emphasis added).[135]

Hence, the AI system seems to act as the epicenter of liability in the French proposal, as it has both the duty to notify the drivers that it is capable of exercising dynamic control – at a certain moment of the trip – and to alert them that it is no longer capable to do so – at another moment of the trip (through a *demand de reprise*).[136]. If we exclude holding the AI system directly liable, this entails imposing obligations and liability (indirectly) on the PCAs, who will have to make sure to place into commerce a vehicle which can fulfill these duties per design.[137]

With regards to the vehicle producer, article L.123-2 proscribes that the producer will be liable for the offences of unintentional harm to the life or integrity of the person[138] committed by the vehicle during periods when the ADS exercised dynamic control of the vehicle, in accordance with its conditions of use, provided that a fault is established within the meaning of Article 121-3 of the French Penal Code.

3.3 *The Joint Report on Automated Vehicles of the UK Law Commissions*

The Joint Report is a 292-page document which contains 75 recommendations on how to develop a new regulatory reform for AVs. It is the result of a four-year work started in 2018 upon request from the UK Government's Centre for Connected and Autonomous Vehicles.

[135] Code de la route, article L.319-3 (emphasis added).

[136] Giuca, *supra* note 129, p. 23.

[137] *ibid*

[138] *Code pénal*, articles 221-6-1, 222-19-1 and 222-20-1.

It represents the first time that the Law Commissions have been asked to develop a legal framework *before* future technological development.[139] The ultimate purpose of the Joint Report is to lead to the adoption of *ad hoc* legislation, ie, the Automated Vehicle Act. As such, it is of utmost interest, since it provides an example of how governments might attempt to regulate negligence failures via *hard* law.

The Joint Report defines an AV as a vehicle that is designed to be capable of driving itself.[140] Self-driving vehicles operate in such a way that they do not need to be controlled and monitored by an individual, for at least a portion of a journey. The drafters of the Report expressly distance themselves from the nomenclature developed by the Society of Automotive Engineers (SAE), which identifies six levels of automation.[141] What is more, the UK Commissions chose to use the term ''self-driving'', which is explicitly discarded by the SAE.[142] They did so deliberately as they wanted to connote a *legal,* rather than mechanical, threshold. As we will see, once this threshold is satisfied, the human in the driving seat (the so-called ''user-in-charge'') will no longer be liable for the (damages caused by) the dynamic driving task.[143]

As a matter of fact, choosing to discard the SAE taxonomy might supersede criticisms regarding a lack of clarity on the difference between SAE levels 2 (driver assistance) and 3 (''eyes off the road'').[144] As stated by Chesterman, ''level three … marks an inflection point''

[139] LCR, *supra* note 82, [1.1]. Although the Joint Report touches upon a large variety of topics, we will focus exclusively on those recommendations regarding the allocation of criminal liability when the control of the vehicle is shared between the ADS and a human being.

[140] *ibid* [1.10].

[141] They are: 1 – no automation; 2 – driver assistance; 3 – partial automation; 4 – conditional automation; 5- high automation; 6 – full automation. Society for Automotive Engineers International (SAE), J3016 Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles (April 2021) (''SAE Taxonomy'').

[142] LCR, *supra* note 82, [7.1].

[143] *ibid* [2.22].

[144] The automated driving feature can perform all the driving tasks but the human in the driving seat, ie, the ''fallback-ready user'', is expected to be receptive to respond to a request to intervene or to an evident failure of the system, but she is not expected to monitor the driving environment. Chesterman, *supra* note 83, p. 35.

meaning that "the driving system is responsible for monitoring the environment and controlling the vehicle".[145] Yet, whether "the importance of that inflection point … is apparent when it comes to liability, though where level two ends and level three begins may not always be clear".[146] Chesterman uses the (in)famous Elaine Herzberg case as an example to prove his affirmation. Elaine Herzberg was struck and killed by an automated Uber test vehicle transporting a human operator. The car failed to recognize whether Herzberg was a pedestrian, a vehicle, or a bicycle. As reported by the National Transportation Safety Board,[147] the probable cause of the crash was the failure of the vehicle operator to monitor the driving environment and the operation of the ADS because they were visually distracted throughout the trip by their personal cell phone. According to Chesterman, even though the Uber test vehicle was a level 2 one, its driver "appears to have acted as though it were level three",[148] which proves that "[t]hough satisfying the legal fiction that there is a 'driver', the reality is that humans not actively engaged in a task such as driving – that is, when their hands are off the wheel – are unlikely to maintain for any length of time the level of attention necessary to serve the function of backup driver in an emergency".[149]

The Herzberg case appears paradigmatic of what is often referred to as *automation bias* or *automation complacency*, a well-known phenomenon in the field of aviation. It refers to the state of a monitoring human experiencing a "low index of suspicion".[150] In other words, "[w]hen you automate any part of a task, the human overseer starts to trust that the machine has it handled and stops paying attention".[151] Automation bias is particularly common when the

[145] *ibid*

[146] *ibid*

[147] National Transportation Safety Board, "Highway Accident Report. Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian" (19 November 2019), NTSB/HAR-19/03 PB2019-101402.

[148] Chesterman, *supra* note 83, p. 35

[149] *ibid*

[150] Raja Parasuraman and Dietrich H. Manzey, "Complacency and Bias in Human Use of Automation: An Attentional Integration" (2010), 52 *Human Factors The Journal of the Human Factors and Ergonomics Society*, Issue 3, 382 quoting E.L. Wiener, "Complacency: Is the term useful for air safety?" (1981), *Proceedings of the 26th Corporate Aviation Safety Seminar*, p. 119.

[151] Lauren Smiley, "I'm the Operator": The Aftermath of a Self-Driving Tragedy, *Wired* (8 March 2022) <https://www.wired.com/story/uber-self-driving-car-fatal-crash>

automated system is highly, but not *perfectly*, reliable, and occurs even if the operator was informed that the system is not perfect.[152] In a pure[153] automation bias incident, no fault actually occurs from the part of the system, nor from the human-system interface structure. The system correctly cedes back control to the human as intended by its designers, and there is enough "control" (in the sense of actual ability of the human to intervene, contrary to the problems discussed above[154]) for the operator to intervene, but a psychological default causes the human to fail in this task. As such, automation bias can be viewed as a (human) negligence failure, instead of one attributable to the machine or its design. Does the scheme proposed by the UK Law Commissions address the concerns raised by automation bias and the lack of "de facto" control? These questions will guide us in the following analysis.

3.3.1 *The Authorisation Scheme*

First of all, the UK Law Commissions propose the introduction of a new and independent authorisation scheme[155] to evaluate whether an ADS feature can be considered as self-driving according to the law or not.[156] An ADS "feature" is defined as "a combination of software and hardware which allows a vehicle to drive itself in a particular operational design domain (such as a motorway)".[157] The authorization would entail that, once the ADS feature is correctly engaged,

---

Footnote 151 continued
accessed 27 July 2022. https://www.wired.com/story/uber-self-driving-car-fatal-crash/

[152] Parasuraman and others, *supra* note 67, p. 290.

[153] Often, however, automation bias cases are not "purely" caused by the operator's negligent behaviour, and the operator's complacency was itself a product external factors such workplace culture or overly optimistic guarantees that the system "can be trusted". See eg US Department of Defense, "Investigation Report: Formal Investigation into the Circumstances Surrounding the Downing of Iran Air Flight 655 on 3 July 1988" (1988) AD-A203 557.

[154] See above Subsection 2.4.

[155] Authorization would consist of a separate procedure from domestic, European, or international approvals, which instead regard whether the vehicle can be placed on the market.

[156] LCR, *supra* note 82, ch.2, pp. 69 et seq.

[157] *ibid* note 239, p. 135. This entails that, in order for this scheme to be applicable in fields other than AVs, there would need to be a uniform definition of when a particular level of automation is engaged. The exact way automation is expressed will differ from application to application.

the human in the driving seat would acquire by law the new role of ''user-in-charge'', causing a change in the allocation of liability, as it will be discussed below.

Each ADS feature would have to be assessed on three different aspects.

First, whether the feature reaches the legal threshold to be labelled as self-driving.[158] The term is so cogent that the drafters recommend that it becomes ''protected'', in the sense of being safeguarded by two specific criminal offences: Offence 1, ''Describing unauthorised driving automation as 'self-driving'''[159] and Offence 2, ''Misleading drivers that a vehicle does not need to be monitored''.[160]

Second, each ADS feature must be able to control the vehicle in a legal and safe way, even if the human user is not monitoring the driving environment, the vehicle, or the way the way the vehicle drives. Safety plays a vital role in the drafted regulation: what should the safety standard be? How should it be established in practice and by who? Indeed, defining such a standard is quite a challenging task: should it be, for example, an amount x of failures for a time t? Or should it be a more qualitative descriptor of the AI's performance? The Law Commission believe that this is a political issue, hence they recommend that the new Automated Vehicle Act ''require the Secretary of State for Transport to publish a safety standard against which the safety of automated driving can be measured''[161] which ''should include a comparison with harm caused by human drivers in Great Britain''.[162]

The evaluation of the safety of AVs shall be done through empiric research: the Joint Report delegates the responsibility of collecting data, which compares the safety of automated and conventional driving to the ''AV in-use regulator'', to a new legal subject: the AV in-use regulator, to be instituted via legislation.[163] Conducting such comparisons, as noted by the Law Commissions, might prove problematic. One of the reasons of this difficulty is that ''road safety statistics provide reliable data about rare events (such as fatalities)

[158] LCR, *supra* note 82, [2.57].

[159] *ibid* [7.21].

[160] *ibid* [7.38].

[161] *ibid* [4.66].

[162] *ibid*

[163] The AV in-use regulator will have several duties, such as applying regulatory sanctions for breaches of traffic rules by an AV driving itself. *ibid* ch. 6.

but less data about more common events, such as minor collisions''.[164] Nevertheless, measuring the performance of the AVs against those of human drivers would ensure public acceptance: ''When deaths and injuries occur, it will be important to reassure the public that AVs are nevertheless safer than human drivers, and to have the evidence to support this claim''.[165] On the one hand, this is judicious for its evidence-based approach and flexibility (as it is likely performance standards will evolve as new models are tested and released), but on the other hand, it also carries some risk: referring such a delicate evaluation to politics could lead to abuse, for example in jurisdictions which are subject to influence of lobbies that do not have victims' interest in mind. Moreover, the standard would have the status of a statutory guidance, meaning that it would not have a binding effect comparable to legislation.

Third, the authorization authority will evaluate whether the ADS entity (ADSE) has sufficient resources to keep the vehicle updated and compliant with traffic laws in Great Britain and to deal with any kind of issue that might arise.

The Law Commissions explicitly state that they aim to ''*draw a bright line*'':[166] criminal liability of the person sitting in the driving seat of a self-driving vehicle shall be excluded for *any* harm arising from the dynamic driving task, in all cases where the offence is committed by a vehicle which was previously authorized to deploy self-driving features, assuming that those features were properly engaged. As was already outlined above, the act of ''drawing a (bright) line'' is a recurrent theme in the regulatory schemes discussed in this research. Perhaps this can be reconnected to the fact that prescribing immunity clauses entails identifying *finite* areas of non-punishment inside *larger* areas of punishment, similar to drawing a Euler diagram. Moreover, by invoking the concept of a ''clear bright line'',[167] the Law Commissions also attribute a strong communicative function to the (new) legal regime: it will separate AI systems ''which require

[164] *ibid* [6.29].

[165] *ibid* [4.62].

[166] *ibid* [5.46]. A dynamic driving task is defined as ''the real-time operational and tactical functions required to operate a vehicle in on-road traffic. It includes steering, accelerating and braking together with object and event detection and response''. *ibid* xviii.

[167] *ibid* [3.5].

attention and those that do not'',[168] liability from non-liability, wrongdoers from welldoers.

3.3.2 *New Legal Actors, Take Two: the British Users-In-Charge*

The recommendations create three new legal actors: the user-in-charge, the Authorised Self-Driving-Entity and the No-User-In-Charge operator.

Starting from the first, the user-in-charge is defined as the human being sitting in the driver's seat while a self-driving feature is engaged. The main role of the user-in-charge is ''to take over driving, either following a transition demand or because of conscious choice''.[169] As already mentioned, users-in-charge enjoy immunity from ''driving offences'',[170] provided that they have engaged the ADS correctly and that they have not tampered with the system.[171] Driving offences do not constitute a pre-existing category of crimes in UK legislation. They are defined in Joint Report as any offence involving ''a breach of duty to monitor the driving environment and respond appropriately by using the vehicle controls to steer, accelerate, brake, turn on lights or indicate''.[172] Examples of dynamic driving offences are dangerous driving, careless driving, and exceeding the speed limit. This is possibly the most ground-breaking change advised by the joint report.

The definition of user-in-charge can be broken up into four characteristics. A user in charge is:

(1) *an individual*, ie, a human or ''natural person'', rather than an organisation;
(2) *who is in the vehicle,* hence not standing nearby or in a remote operations centre;
(3) *in position to operate the driving controls*, which for current vehicle design entails that they are in the driving seat;
(4) *while an ADS feature requiring a user-in-charge is engaged.*[173]

---

[168] *ibid*

[169] *ibid* [4.1].

[170] The user-in-charge remains liable for non-dynamic offences and is responsible for other aspects connected to driving such as the duties to carry insurance and to ensure that child passengers wear seatbelts. cf. *ibid* [8.103].

[171] *ibid* [8.79].

[172] *ibid* [8.62].

[173] An ADS feature is engaged when it is switched on and remains so until the individual takes control of the vehicle, the transition period ends, or it switches off at the end of a journey. Law Commissions, ''Automated Vehicles: Summary of joint report'', HC 1068 SG/2022/15, 26 January 2022 [4.2].

The user-in-charge is no average (reasonable) agent. Certainly, users-in-charge should be deemed criminally liable for being unqualified or unfit to drive, much like "average drivers" who are liable for acts such as unlicensed driving or driving under the influence of substances. Yet, there is more: the user-in-charge must not only be "qualified and fit to drive", but also "receptive to a transition demand" and comply with other "driver responsibilities", which include insuring the vehicle and reporting accidents.[174]

The Joint Report distinguishes between the duties of monitoring and of receptivity. As an example, the Law Commissions quote the SAE Taxonomy, according to which "A person who becomes aware of a fire alarm or a telephone ringing may not necessarily have been monitoring the fire alarm or the telephone".[175] *Monitoring* entails checking the driving environment, the vehicle, or the way it drives. An ADS feature can be considered as self-driving only if it excludes this duty. Hence, the user-in-charge is not expected to perform a monitoring task. *Receptivity*, instead, entails being receptive to a transition demand, ie, the request by the vehicle for the human user to take over the dynamic driving: *this* is the duty imparted on the user-in-charge. The transition demand must be communicated by clear, multi-sensory signals and give the user-in-charge sufficient time to gain situational awareness.[176]

The duty of receptivity is also present in the French amendment, which provides that the driver shall constantly be in a condition and in a position to respond to a transition demand from the ADS.[177] Once again, we must emphasize the caveat attached to such *reprise* clauses:[178] this timeframe must allow users-in-charge sufficient opportunity to obtain *de facto* awareness and control to avoid raising the same problems identified in Subsection 2.4. Indeed, when it comes to transition demands and liability, time is of the essence:[179] as soon as the transition period is over, the user-in-charge loses immunity and is legally treated as a driver. Yet, the Law Commissions also note that

[174] LCR, *supra* note 82, [8.6].

[175] SAE Taxonomy, *supra* note 141, p. 12.

[176] LCR, *supra* note 82, [2.15].

[177] *Code de la route*, L. 123-1.

[178] See also Subsection 3.2.

[179] "The length of the period is legally significant", LCR, *supra* note 82, [3.27].

they are not in the position to specify how long this period should be, leaving then a major gap in the proposed regulation.[180]

Furthermore, we need to address an additional specification, enclosed in a previous consultation paper, that was not included in the Joint Report: in order for users to be *receptive*, they need to know *what* they must be receptive to. Furthermore, they might need to "rehearse how to respond appropriately if the stimulus arises". As a matter of fact, "[t]hat is why, in addition to installing fire alarms, organisations have fire drill".[181] How shall a normal driver, then, become a fit user-in-charge? One could argue that it would be reasonable for legislators to provide for a mandatory "special" license (with special fitness tests) for "autonomous car" drivers. This is a very important and desirable clause which we recommend strongly for all similar attempts at legislation. To recall, one of the issues with the epistemic problem we identified was that there may be an *incentive* for users to not understand the system or how to react if this can potentially reduce their chances of criminal conviction.[182] However, adding a license requirement with mandatory training on the AV's workings and how to properly react to a transition demand closes this potential escape route. A user-in-charge could no longer claim ignorance, as this is automatically disproven by the fact that they possess the license which allowed them to operate the AV in the first place.

Moreover, as stated above, the definition of user-in-charge represents a topic on which the UK and Singaporean approaches appear radically different. Let us consider for example the act of taking over the control of the vehicle in order to follow the order of a police officer to stop after an accident. According to the Singaporean approach, this would amount to an instance of behaviour which could be codified by the legislator as a required standard of conduct to fulfil, in order not to be negligent actors.[183] Following the UK approach, it would instead represent an instance of dynamic driving.[184] This entails that, in the former system, users-in-charge would be liable if they did not intervene to stop the car, regardless of whether the car instructed them to do so; in the latter, that it is the AV that should

[180] Ideally, this should be determined through actual data obtained from empiric testing.

[181] Scottish Law Commission, "Automated Vehicles: Consultation Paper 3 – A regulatory framework for automated vehicles. A joint consultation paper" (Law Com No 258), Consultation Paper No 252, [4.27].

[182] See also Subsection 2.2.

[183] SAL, *supra* note 92, [4.27].

[184] LCR, *supra* note 82, [8.68].

either stop or issue a transition demand once it detects an accident. Here, the liability of the users-in-charge for not stopping the car following the policeman's order would only subsist if they failed to respond to the transition demand, assuming ADS feature was built and approved to deliver one in such situations.[185]

We mentioned that the immunity clause for the user-in-charge is lifted if the user fails to respond to a transition demand. In these cases, the ADS "should carry out a sufficient risk mitigation manoeuvre . . . (at a minimum) the vehicle should come to a controlled stop in lane with its hazard lights flashing".[186] Such a provision appears to be sufficiently easy to transpose to other domains than AV. Governments could demand that producers must program their AI system as to be able to take an action that maximally reduces the risk of unwanted consequences. But what would happen to users-in-charge (now drivers) in terms of liability if they fail to take over? The recommendations only state that "the law should impose consequences", without providing any kind of further instructions.[187] Again, a *lacuna* occurs.

To conclude, let us now bridge back to French approach. According to the UK Commissions, the 2021 *ordonnance* is too simplistic when it deals with the "dynamic/non-dynamic divide",[188] since it defines dynamic driving offences merely as those which derive from a vehicle manoeuvre, when driving is delegated to an ADS. The Joint Report identifies two major differences between the French and the UK approach. First, the fact that the French model requires the driver to be responsive to some events, such as the presence of emergency vehicles on the road, while the UK model does not. Second, the fact that according to the French model the immunity clause is triggered only if the driver has engaged the ADS in compliance with its terms of use, where instead the UK model strongly discards this option, arguing that it would be unrealistic for users to "check detailed lists of terms of use before engaging an ADS".[189] The proposed solution is one based on the principle of "safety by design": the ADS should be programmed to as to not operate outside its operational design domain.[190] Indeed, by doing so, the UK Commissions take a strong stand against the risks of driver-scapegoating.

[185] *ibid* [8.90].

[186] *ibid* [3.42].

[187] *ibid* [8.132].

[188] *ibid* [8.10].

[189] *ibid* [8.63–8.75].

[190] *ibid*

3.3.3 *ASDEs, NUIC-Operators, & Duty of Candour*

As a final point, it is relevant to focus briefly on the other two new legal subjects which are introduced by the Joint Report: the Authorised Self-Driving-Entity (ADSE) and the No-User-In-Charge (NUIC) operators. These subjects are legal persons, rather than natural persons, and might coincide in cases where the vehicle manufacturer or developer is also the one providing a passenger service.

An ASDE is defined as "the entity that puts an AV forward for authorisation as having self-driving features. It may be the vehicle manufacturer, or a software designer, or a joint venture between the two".[191] The ASDE will have the duty to prove in the authorisation process that the user-in-charge has sufficient time to gain situational awareness in cases of a transition demand and, if they fail to respond, that the vehicle is capable of sufficient mitigation against the risk of a crash. Other duties of the ASDE include those connected to safety (such as ensuring that the vehicle continues to drive safely and in accordance with road rules) and duties of disclosure.[192]

The NUIC operator is intended as a licensed legal person which oversees vehicles possessing a NUIC feature. While on a vehicle deploying NUIC features, a whole journey can be completed without any kind of intervention by a human on board. This does not mean that there would be no human being on board, but that when the ADS feature is of NUIC nature, any human in the car will be considered as a mere passenger. The NUIC operators need to have "oversight" of the vehicle any time a NUIC feature is engaged on a road or in another public place: they are "expected to respond to alerts from the vehicle if it encounters a problem it cannot deal with, or if it is involved in a collision", but they are not expected to monitor the driving environment.[193] Oversight duties would include both remote assistance (for example, if the ADS detects an object in its lane which is too large to avoid and stops, remote assistance could imply providing instructions to the vehicle on how to deal with the obstruction) and fleet operations (for example, dealing with law-enforcement agencies or paying tolls).[194]

The UK commissions believe that the NUIC operator should not be the addressee of statutory immunity from criminal offences such as the

[191] *ibid* [2.41].

[192] *ibid* [5.65–5.96].

[193] *ibid* [2.48].

[194] *ibid* [9.14].

one proscribed for the user-in-charge.[195] As stated in Recommendation 56, the regulator shall have powers to impose only *regulatory* sanctions (such as warnings, civil penalties, suspension or withdrawal of licence) upon NUIC operators.[196] Moreover, certain offences regarding the "use" of the vehicle might apply to a NUIC operator, depending on whether the NUIC operator is the registered keeper or the owner of the vehicle. Additionally, the individual staff of the NUIC involved in remote driving of the vehicle could face the same criminal liabilities as drivers, for example if they are not trained or qualified enough.[197] Yet, the Joint Report does not advise in favour on introducing new criminal offences relating to individual assistants.

Finally, the Joint Report recommends the introduction of five new offences for ASDE and NUIC operators which are related to violations of a "duty of candour":[198] Offences 1 and 2 punish the non-disclosure or misrepresentations of information to the regulator; Offence 3 punishes non-disclosure and misrepresentations in responding to regulators' requests; Offence 4 punishes the consensual or conniving conduct of senior managers of ASDEs or NUIC operators in cases where the ASDE/NUIC operator has committed Offence 1, 2 or 3; Offence 5 would punish offences committed by the nominated person, ie, the person who signed the relevant safety case or response to the request for information, in cases where the ASDE/NUIC operator has committed offence 1, 2 or 3. The sixth Offence, instead, aggravates Offences 1 to 5 in cases where the misrepresentation or non-disclosure leads to death or serious injury of a subject. These novel offences are relatively effective responses to the problem related to distance and many hands of PCAs, as the primary challenge attached to PCAs is linking them with a specific offence occurring at vast temporal and physical distances from their contribution in the AV life-cycle.[199] Introducing specific Offences 1-6 mitigates this problem by opening the possibility to prosecute them for a separate offence that is much more closely linked to their roles in the production and distribution process.

[195] *ibid* [9.132].

[196] *ibid* [9.120].

[197] *ibid* [9.42].

[198] *ibid* chapter 11.

[199] See also Subsection 2.5.

IV CONCLUSIONS

AVs are promising technologies which have the potential to greatly improve societal welfare and reduce unneeded harm arising from human error. However, like all tools, they are not perfect: they will fail, and sometimes such failures may have catastrophic consequences. A reasonable society must not only embrace the advantages of such technologies, but also ensure that an effective and sufficiently tailored legal regime is adopted to safeguard the rights of both victims and potential accused. Such a legal regime must thread an admittedly delicate balance between two undesirable extremes: ignoring the specificities of the technology altogether while scapegoating persons who had no knowledge or control of the harmful outcome, or being too permissive of these changes while excluding any possible criminal liability altogether.

In this light, in Section II we first outlined the diverse reasons why AI systems are, indeed, different. A major aspect is their ML component, which often makes the AI's workings no longer susceptible to intuition. Even designers frequently cannot exactly predict or understand how their ML systems work, and this will apply to an exponential degree for drivers of AVs who have no background in AI. Together with the problem of generic risk, these characteristics can make the epistemic element required for *mens rea* – knowledge – functionally absent. In addition, we also highlighted that *de facto* control is crucial for criminal liability: only acts or omissions over which the accused actually had control can be attributed to them. We have seen that this problem is particularly prescient for so-called *reprise* clauses, which may not always take into consideration whether the users had the functional capacity to intervene properly. Finally, focalizing guilt unto specific PCAs is made difficult from the sheer magnitude of actors and interactions involved, which we referred to as the many hands problem. We argued that the characteristics of modern AI bring about specific circumstances (such as the epistemic problem and the problem of many hands) that can lead to a malfunction of the assumptions underlying the criminal legal concept of negligence. We refer to these malfunctions of the attribution mechanism as "negligence failures".

Following from these premises, we scrutinized three diverse approaches to fix negligence failures: the Singaporean proposals on a general criminal liability framework for AI offences, the French amendments to the Road Act, and the UK proposal on criminal lia-

bility arising from AVs. The relevance of these initiatives goes beyond their territorial scope of application, since they provide a perfect study sample of how governments can regulate this matter in the future.

We found that the specimens share some common characteristic. To begin with, they intend to ''draw a bright line'': assuming that, as with any other technology, we won't be able to reduce the risks of AI harm to zero, they aim to distinguish in this area of risk zones of legality from zones of illegality. Some do so more clearly than others. Furthermore, they introduce a new legal vocabulary, which comprises of new legal subjects, such as the ''user-in-charge''. This is an instance of a general trend in AI-regulation, according to which AI technology is regarded as ''something new'', hence calling for new legal constructs.

Let us conclude with a few considerations on future policy initiatives. First, we recommend that new regulations on complex AI-based tools establish a license regime with mandatory training for the user. Indeed, such training should be comprehensive enough to avoid situations where the accused might seek to avoid liability by arguing that the AI system was simply inscrutable and that, as a consequence, they lacked the *knowledge* of risks attached to its use or that they lacked the *required skills and reflex* to properly intervene when required. Second, *reprise* clauses must allow *de facto* control. As discussed above, setting these safety standards is relevant for our discussion, since their violation is often connected to establishing negligent liability. Thus, an evidence-based approach should be taken, drawing from empirical data and human-machine interaction theory to ensure that no scapegoating of either the user or PCAs occurs. We acknowledge that in this area drawing a *clear* bright line might be very challenging. For example, in the field of AVs, it might be troublesome to identify the *precise* number of seconds which are needed for a user-in-command to react upon a takeover request. In any case, *more* empiric data and technical knowledge will surely enter future criminal courtrooms, where decision-making authorities will be tasked with applying the legal frameworks outlined above to real-life scenarios and to real-life users-in-charge.

While we used AVs as an illustrative case-study in this article to analyse these ''negligence fixes'', we discussed the technology at a sufficient level of abstraction to allow transposition of these conclusions to any domain where contemporary, ML-dominated AI is utilised.[200] In this respect, future discussion will need to focus on the

[200] See *supra* note 12.

possibility and efficacy of adapting similar regimes to address the ''negligence failures'' in those domains.

DECLARATIONS

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical Approval They authors further declare that they have complied with ethical standards as established by the Committee on Publication Ethics (COPE).