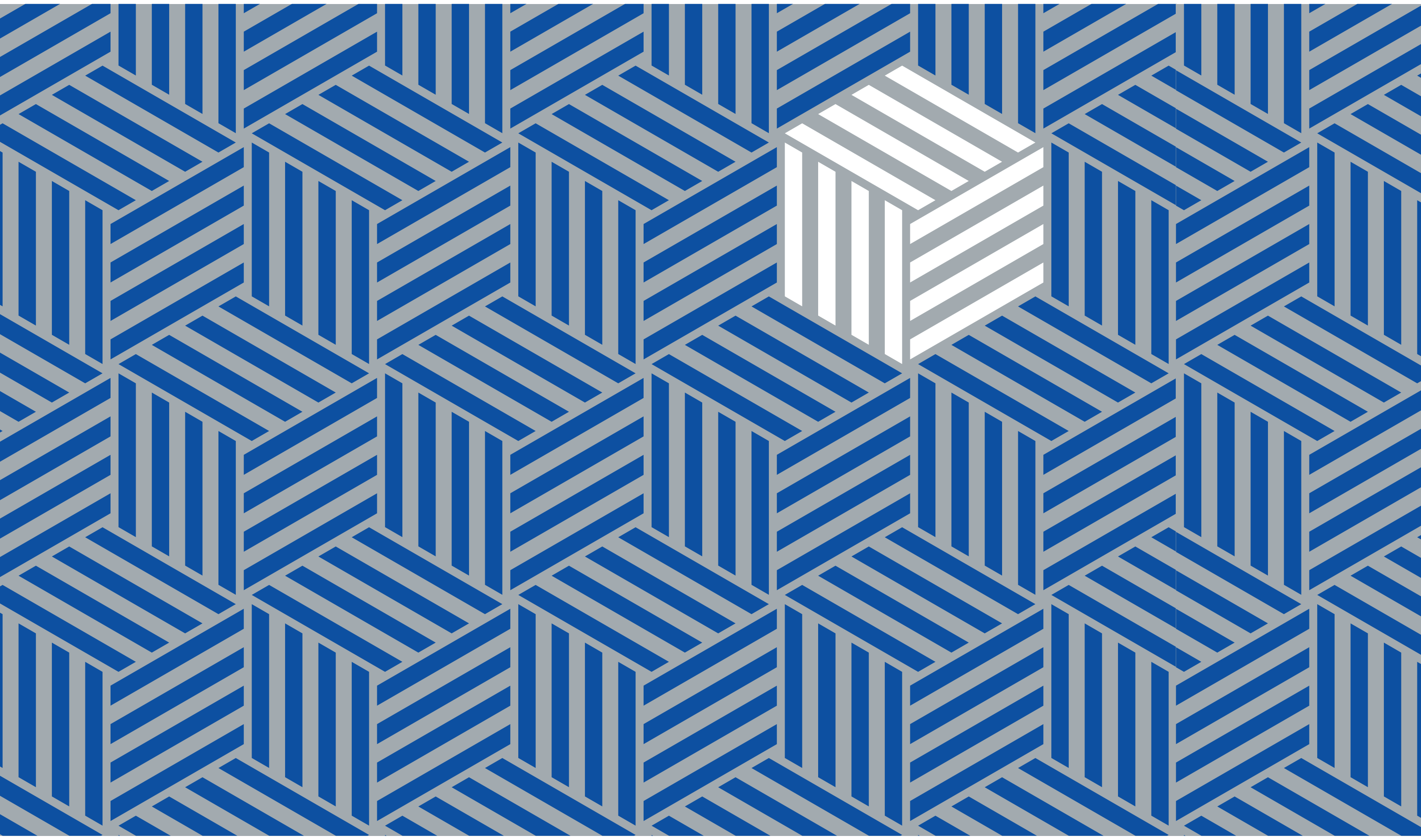
Singapore Academy of Law
Law Reform Committee

Report on Criminal Liability, Robotics and AI Systems

February 2021



Singapore Academy of Law
Law Reform Committee

Report on Criminal Liability, Robotics and AI Systems

February 2021

Part of the *Impact of Robotics and Artificial Intelligence on the Law* series

SINGAPORE
ACADEMY
OF LAW

COPYRIGHT NOTICE

Members of the Robotics and Artificial Intelligence Subcommittee

1. The Honourable Justice Kannan Ramesh (co-chair)
2. Charles Lim Aeng Cheng (co-chair)
3. Chen Siyuan
4. Desmond Chew
5. Josh Lee Kok Thong
6. Gilbert Leong
7. Beverly Lim
8. Sampson Lim
9. Ronald Wong Jian Jie
10. Yvonne Tan Peck Hong
11. Yeong Zee Kin

The report was edited by Simon Constantine, Deputy Research Director, Singapore Academy of Law.

An electronic copy of this report may be accessed from the Singapore Academy of Law website at *https://www.sal.org.sg/Resources-Tools/Law-Reform/Law-Reform-e-Archive*.

National Library Board, Singapore Cataloguing in Publication Data

Names: Singapore Academy of Law. Law Reform Committee. | Constantine, Simon, editor.
Title: Report on criminal liability, robotics and AI systems / Singapore Academy of Law, Law Reform Committee ; edited by Simon Constantine.
Description: Singapore : Singapore Academy of Law, [2021]
Identifiers: OCN 1235852527 | ISBN 978-981-18-0164-8 (softcover) | ISBN 978-981-18-0165-5 (e-book)
Subjects: LCSH: Artificial intelligence–Law and legislation–Singapore–Criminal provisions. | Robotics–Law and legislation–Singapore–Criminal provisions. | Criminal liability.
Classification: DDC 345.595704–dc23

ISBN 978-981-18-0164-8 (softcover)
978-981-18-0165-5 (e-book)

About the Law Reform Committee

The Law Reform Committee ("LRC") of the Singapore Academy of Law makes recommendations to the authorities on the need for legislation in any particular area or subject of the law. In addition, the Committee reviews any legislation before Parliament and makes recommendations for amendments to legislation (if any) and for carrying out law reform.

Comments and feedback on this report should be addressed to:

Law Reform Committee
Attn: Law Reform Director
Singapore Academy of Law
1 Coleman Street
#08-06 The Adelphi
Singapore 179803
Tel: +65 6332 4070
Fax: +65 6333 9747
Email: *lawreform@sal.org.sg*

IMPACT OF ROBOTICS AND ARTIFICIAL INTELLIGENCE ON THE LAW

SERIES PREFACE

It has been said that we are at an inflection point in the development and use of Artificial Intelligence (AI). The exponential growth in data in the past decade – from 2 trillion gigabytes in 2010 to around 33 trillion at the end of 2018, and an anticipated 175 trillion by 2025 – has enabled giant datasets to be compiled and used as the basis for developing ever-more sophisticated AI systems.

Those systems are in turn being used – in commercial, military, consumer and other contexts – to enhance humans' ability to carry out tasks, or to replace humans altogether. From self-driving cars and robotic carers, to autonomous weapons and automated financial trading systems, robotic and other data-driven AI systems are increasingly becoming the cornerstones of our economies and our daily lives. Increased automation promises significant societal benefits. Yet, as ever more processes are carried out without the involvement of a 'human actor', the focus turns to how those robots and other autonomous systems operate, how they 'learn', and the data on which they base their decisions to act.

Even in Singapore, which ranks among the world's leading nations in the International Development Research Centre's Government Artificial Intelligence Readiness Index, questions inevitably arise as to whether existing systems of law, regulation and wider public policy remain 'fit for purpose', given the pace and ceaselessness of change. That is, do they encourage and enable innovation, economic growth and public welfare, while at the same time offering protection against misuse and physical, financial or psychological harm to individuals?

To this end, the Singapore Academy of Law's Law Reform Committee ('LRC') established a Subcommittee on Robotics and Artificial Intelligence to consider, and make recommendations regarding, the application of the law to AI systems.

Having considered current Singapore law, as well as legal and policy developments in other parts of the world, the LRC is now publishing a series of reports addressing discrete legal issues arising in an AI context.

There is currently much work being undertaken at a national and international level in this field. Domestically, the Singapore Government has published the second edition of its Model AI Governance Framework and launched a National Artificial Intelligence Strategy to reap the benefits of systematic and extensive application of new technologies. The LRC hopes that its reports will complement and contribute to these efforts and help Singapore law – through legislation or 'soft law' – to develop in a way that fosters socially and economically beneficial development and use of robotic and AI-driven technologies.

The series does not purport to offer comprehensive solutions to the many issues raised. The LRC hopes, however, that it will stimulate systematic thought and debate on these issues by policy makers, legislators, industry, the legal profession and the public.

OTHER REPORTS IN THIS SERIES

- Applying Ethical Principles for Artificial Intelligence and Autonomous Systems in Regulatory Reform (published July 2020)
- Rethinking Database Rights and Data Ownership in an AI World (published July 2020)
- Report on the Attribution of Civil Liability for Accidents Involving Autonomous Cars (published September 2020)

TABLE OF CONTENTS

REPORT ON CRIMINAL LIABILITY, ROBOTICS AND AI SYSTEMS

EXECUTIVE SUMMARY

1 Autonomous robotic and artificial intelligence (‘**RAI**’) systems are increasingly being deployed across all aspects of our lives, and engaging with humans ever-more frequently. Alongside the numerous benefits likely to result from such developments, the risk that those systems cause physical, emotional or economic harm or damage to humans or property can also be expected to increase.

2 Against that backdrop, this report highlights certain potential risks posed to humans and property by the use of RAI systems, and examines whether and how Singapore’s criminal laws may apply, and criminal liability may arise, in such situations.

A CRIMINAL LIABILITY AND RAI SYSTEMS

3 Attribution of criminal liability to a person generally requires both a wrongful act (or, in certain cases, omission) and – strict liability offences aside – a mental or “fault” element on the part of the person carrying out the act. That fault element (also known as “*mens rea*”) may involve intention, wilfulness, knowledge, rashness, or negligence.

4 The increasing autonomy of RAI systems, and the corresponding reduction in the roles of humans in the actions of those systems, can raise challenges in attributing criminal liability and holding someone responsible where harm is caused.

- While criminal liability can be imposed on natural or legal persons – and thus on both humans and corporate entities – an RAI system is not a legal person on which criminal responsibility could be placed directly.
- Moreover, every decision made by an RAI system is the result of a long causation chain. The ultimate decision to act in a certain way – which in a non-RAI context would be made only by the human undertaking that act – is instead distributed further up the decision-making chain, muddying the identification of blameworthy actors.

5 Complex questions therefore arise as to (a) which aspect of the RAI system factually caused it to act the way it did (resulting in harm), (b) which party (or parties) – be that the system manufacturer, the system owner, a component manufacturer, or a software developer – was responsible for that aspect, and (c) whether that party could have foreseen or mitigated the harm. Determining those issues may entail a highly

complex assessment (factually and technologically), which may not yield a clear outcome.

6 More broadly, consideration must be given to whether – from a policy perspective – it is appropriate to apply *criminal* law in a given context. This involves balancing societal imperatives (e.g., encouraging those who deploy RAI systems to take necessary safety measures; setting socially-acceptable standards for use of RAI systems; condemning morally unacceptable behaviour) with the desire to avoid having a chilling effect on innovation and RAI system development through the imposition of unreasonable burdens or disproportionate liability exposure.

7 Given the wide variety of both RAI technologies and the applications and settings in which they may be used (each entailing differing sources and levels of risk, responsibility and potential benefit), a 'one size fits all' approach to questions of liability is not practicable. In some circumstances, the seriousness of the (potential) harm, the degree of moral culpability or the need for deterrence may justify making certain wrongful acts subject to criminal penalties. In others, non-criminal enforcement and penalties may be deemed a sufficient or preferable sanction and deterrent against future harms.

8 In light of the above, whether – and on whom – any criminal liability for a harmful act involving an RAI system should be imposed is likely to be a function of: (a) the severity and risk of actual or potential harm inherent in the use of the system in the relevant context; (b) the level of automation of that system; and (c) the degree of human oversight over, and involvement in, the system's decision-making (if any).

- Where the level of automation is more limited and/or the degree of human oversight greater, it is likely to be relatively straightforward to identify a human user to whom liability may (where appropriate) be attributed or on whom certain obligations may be imposed.
- In some instances, particularly as automation increases, it may be necessary or appropriate to define expressly in laws or regulations (a) who the user of the RAI system is deemed to be, and/or (b) the responsibilities on them to oversee or control its behaviours.
- With fully-automated, "human-out-of-the-loop" systems (where the RAI system, within set parameters, makes and executes decisions without any human input or interaction), there may be no identifiable human user involved. This raises questions as to who, if anyone, should be held responsible for any harm caused, and on what basis.

9 Where a human 'user-in-charge' does remain involved in the use or operation of an RAI system, it is useful to distinguish between cases of

intentional criminal use of (or interference with) the RAI system, and those where *non-intentional* criminal harm is caused.

10 <u>Intentional harm</u> – It seems uncontroversial that where a person (including a third party who is not the user-in-charge) *intentionally* uses or causes an RAI system to cause harm, that person should be liable for the act.

- While existing laws (for example, certain offences under the Computer Misuse Act or Penal Code) may be capable of addressing certain malicious uses of RAI systems, there are potential limitations or ambiguities in such coverage.
- As such, we consider that there is merit in reviewing those existing laws to identify and fill any gaps, so as to ensure that intentional harms caused through the use of RAI systems are effectively caught.

11 <u>Non-intentional harm</u> – The operation of RAI systems may result in harm, even though the human user-in-charge did not intend it and there was, for example, no unlawful interference on the part of someone other than the user-in-charge.

- In such a case, it would seem difficult to argue that the human user had the requisite mental state to be found to have committed a criminal offence, at least for offences where the fault element requires intention or knowledge. Nor is it clear that any other natural or legal person could be shown to have intended or known the harm would occur.
- However, the fault element of certain offences can also be satisfied (and thus criminal liability imposed) where a person did not intend the harm, but was criminally negligent. Sections 304A and 338 of the Penal Code, for example, impose criminal liability on a person who causes death or grievous hurt by his or her *rash or negligent* act.
- The Penal Code (s 26F) defines a person doing an act "negligently" for the purposes of any criminal offence as a person omitting to do an act that a reasonable person would do, or doing an act that a reasonable person would not do.

12 It is therefore prudent to consider whether criminal negligence frameworks may provide an appropriate basis for attributing liability where non-intentional harms arise from the operation of RAI systems.

B CRIMINAL NEGLIGENCE

13 Certain harms from RAI systems may fall within the existing negligence-based offences in the Penal Code. In other cases, the RAI system or type of harm (particularly non-physical harm) in question may not fall comfortably within those Penal Code provisions, as presently defined.

14 Moreover, inherent in the way the negligence-based Penal Code offences are framed is the need to both (a) determine what an objective 'reasonable person' would do in a given circumstance, and (b) prove that that standard was breached in the case at hand. Both those aspects can create challenges where applied to harms resulting from the operation of RAI systems.

15 Under the negligence-based offences in the Penal Code, it is for the courts to determine what constitutes 'reasonable' conduct in an individual case, applying or adapting standards from past cases, or defining new ones. This enables the law to adapt to new circumstances and technologies. But it is also inherently uncertain. The concern is that such uncertainty may have a chilling effect on the development or use of RAI systems, if developers or users fear that they may (inadvertently) fall foul of the law.

16 To mitigate such concerns, the nature and extent of the standard of reasonable conduct imposed by the law could be set out more precisely in legislation. However, it is unlikely to be practicable to set such a standard (or standards) in law across the board for all possible applications of RAI systems, all sectors, and/or all possible risks of harm. This is especially so in sectors where RAI technologies and their applications are evolving rapidly.

17 Even where the nature and extent of the standard of conduct can be established, it must still be proved in a particular case that that standard has been *breached*.

18 Where the standard is imposed on a human, that may be relatively straightforward to demonstrate. However, greater challenges arise where harm is caused by the way the RAI system operates but the human user-in-charge was not negligent, or there was no human user at all.

19 In such cases, there may be various reasons, unrelated to the user's actions, why the harm in question arose or why the RAI system otherwise failed to comply with the law. Again, in certain circumstances – such as third party interference – those reasons will be apparent and the person 'to blame' easily identified. However, in others, the reason for the RAI system's actions, and thus what caused the harm, may not be so apparent. For example, the cause of the harm might lie in the RAI's algorithmic software. Particularly with more complex RAI systems, it may be very difficult (in some instances, practically impossible) to establish definitively the process by which the RAI system determined to take a particular action.

20 And even if that can be established, questions then arise as to who was responsible for that aspect of the RAI systems' decision making, and whether they acted in a way which would justify the imposition of criminal liability.

C POSSIBLE ALTERNATIVE APPROACHES

21 There is therefore merit in considering whether other mechanisms for applying criminal law may be preferable to, or could usefully supplement, reliance on criminal negligence. This report considers three such approaches. These are not intended to be exhaustive, nor mutually exclusive.

22 This report does not purport to advocate that any particular basis should be adopted by legislators in relation to a specific form of harm or RAI application. Rather, it aims to explore different ways in which criminal laws might – in principle – be formulated and applied to non-intentional harms 'caused' by RAI systems, and the extent to which those approaches might address the various challenges in imposing criminal liability identified above.

Legal personality for RAI systems

23 One possibility that has been debated is the creation of a new form of legal personality (or 'personhood') for RAI systems, such that criminal liability could be imposed directly on the RAI system itself.

24 This has been compared to the way in which corporations have been accorded legal personality and may be found criminally liable. And on its face, there is some appeal in the possibility that it could help limit the need to get 'under the bonnet' of the RAI system and identify which specific part(s) of that system caused the decision to act as it did, and which of the parties involved in the system's development or deployment should therefore be held responsible for that act.

25 The contrary (and – at the current state of technological advancement – in our view more compelling) argument is that criminal laws should continue to be formulated on the basis that they seek primarily to shape or impact human behaviour. It is somewhat unclear, for example, how imposing criminal liability and sanctions on an RAI system directly would 'punish', 'deter' or 'rehabilitate' *the system itself*. And if the objective is instead to deter or penalise those *responsible for* the RAI system, that could arguably equally be achieved through legal mechanisms that do not require new forms of legal personality to be created.

New offences for computer programs considered by the PCRC

26 An alternative approach was considered by the Penal Code Review Committee (PCRC) in 2018. Specifically, the PCRC put forward (albeit simply as a "starting point for future discussions") two possible new criminal provisions that would, respectively:

- target the creation of risk by developers or operators of computer programs through their rash or negligent creation, alteration or use of a computer program, even where no hurt or injury were caused; and,

- impose a duty on those with control over a computer program to take reasonable steps to cease harms that may result from computer programs after they manifest.

27 Such laws may help address two of the challenges with attributing criminal liability raised above, including potentially in scenarios where the RAI system is acting fully autonomously without direct human involvement or oversight. That is, they aid identification of (a) the (legal) person(s) to whom liability should be attributed, and (b) the parameters of the duties to which that person is subject.

28 However, the PCRC's offences do not stipulate the exact contours of the standards or obligations they impose. As such, they do not necessarily address the difficulty of determining if those standards have been breached (i.e., what constitutes a rash or negligent act or omission in a given case).

29 Furthermore, there still remains the broader policy concern that potential criminal liability for non-intentional harms could have an unintentional chilling effect on innovation and the deployment of societally-beneficial RAI systems in Singapore.

30 Even if an approach akin to that envisaged by the PCRC were to be adopted, therefore, it would appear prudent to limit any such offences to, for example, specific high-risk use cases, and to set out more precisely in legislation the relevant obligations imposed and/or standards to be met in a particular case.

Workplace safety legislation as a model

31 However, there remains the possibility that the operation of an RAI system results in death, serious personal injury or widespread public harm, but no individual can be identified as having directly caused that harm (whether intentionally or negligently).

32 By definition, a negligence-based framework for criminal liability would be inapplicable in such cases, regardless of how precisely the relevant obligations or standards of conduct were set out. Yet there may still be strong public demands for someone to be held accountable. To narrow risks of such an "accountability gap", one approach would be to adopt a model more akin to that in existing workplace safety legislation, where duties are imposed on specified entities to take, so far as is reasonably practicable, such measures as are necessary to avoid harm.

33 In the workplace context, those duties are imposed on occupiers and employers. For RAI systems, it might be whichever entity(ies) is best placed – based on their proximity to the RAI system and its operation, and

their resources – to take action (i.e., to prevent, address and rectify dangers posed by RAI systems) and to change future outcomes.

34 This would shift the focus away from investigators having to determine the specific cause of the harm or prove negligence on the part of a (natural or legal) person, and move instead towards a focus on whether the relevant entity breached its statutory duty to take all reasonably practicable measures to avoid the harm. From an enforcement perspective, this has the advantage of avoiding the need to prove a direct or scientifically precise causation between the harm that resulted and a particular breach of duty (which, as noted above, may be factually and technologically challenging).

35 On the other hand, such statutory duties place a significant onus on defendants, and caution is therefore merited. Nonetheless, for policy reasons, such a burden may be considered to be justified in specific (and likely exceptional) circumstances, or for particular technologies, where:

(a) risks of serious harm are particularly acute or there is felt to be a particular moral imperative on the entity in question to prevent the RAI system causing harm; and,

(b) there is a strong public desire for accountability.

36 Concerns of 'over-criminalisation' could be further mitigated by, for example, calibrating and constraining the sanctions imposed to ensure they are proportionate to the nature of the offence and the entity's degree of blameworthiness (indeed, this is true whatever approach to imposing liability is taken).

37 Ultimately, whether and when it is justified to place such an onus on those responsible for RAI systems is a policy judgment for lawmakers, balancing demands for accountability with the desire not to unduly stifle innovation and impede the societally-beneficial development and use of RAI systems.

38 Even if criminal liability is not considered appropriate in a given instance, models akin to those discussed above might still be suitable for adoption, but reframed so as to impose only regulatory controls and sanctions.

39 Regardless of the approach taken, it is apparent that the use of RAI technologies will continue to give rise to new forms of harm and thus to continue to challenge existing laws and regulations, requiring legislators to respond with agility to new and emergent risks. We hope that this report will be a catalyst to proactive analysis and well-informed debate, even as the technology rapidly evolves.

CHAPTER 1

INTRODUCTION[1]

1.1 Autonomous robotic[2] and artificial intelligence (‘**RAI**’) systems are increasingly being deployed across all aspects of our daily lives, and engaging with humans ever-more frequently. As such deployment and use becomes more widespread and the degree of interaction between RAI systems and humans increases, alongside numerous benefits, the risk that those systems may cause physical, emotional or economic harm or damage to humans or property can also be expected to increase. In recent years, for example, there have already been incidences of:

(a) A self-driving car colliding with and killing a pedestrian, having initially misidentified her first as an unknown object and then as a vehicle;[3, 4]

(b) AI-based software being used to impersonate the voice of the chief executive of a multinational company to deceive another employee into making a fraudulent transfer of €220,000 (S$350,000);[5]

[1] The subcommittee wishes to express its gratitude to Ms. Geraldine Zhang from the Singapore Management University and Mr. Zane Chong Weng Teng from the National University of Singapore for their assistance in and contributions to the research and writing of this report.

[2] Although the term ‘robotics’ can be used to describe any computer-controlled machine used to perform tasks, this report is concerned only with those robotic technologies powered by an artificial intelligence (AI) component, in particular those where the AI system in question is able to act (wholly or partly) autonomously.

[3] National Transport Safety Board, Accident Report HAR1903: *Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian – Tempe, Arizona, March 18, 2018* (19 November 2019) <*https://www.ntsb.gov/investigations/AccidentReports/Pages/HAR1903.aspx*> (accessed 1 February 2021).

[4] Autonomous vehicles have also been involved in accidents (fortunately, non-fatal) in Singapore. In 2017, a Mass Rapid Transit passenger train in automatic driving mode collided with a stationary train, injuring 38 people. Investigations found no human error; rather the unexpected disabling of a protective feature on the stationary train caused the other train to recognise it as a three-car, not a six-car, train. Adrian Lim, “Joo Koon collision: ‘Inadvertent removal’ of software fix led to collision”, Straits Times (16 November 2017) <*https://www.straitstimes.com/singapore/transport/inadvertent-removal-of-software-fix-led-to-collision*> (accessed 1 February 2021). Such driverless public transportation is also presently being expanded in Singapore, in particular to public buses in certain locations, raising new issues regarding how, and by whom, passenger safety should be ensured. See Natalie Tan & Clement Yong, “Driverless bus trials draw 320, including curious passengers”, Straits Times (1 February 2021) <*https://www.straitstimes.com/singapore/transport/driverless-bus-trials-draw-320-including-curious-passengers*> (accessed 1 February 2021).

[5] Catherine Stupp, “Fraudsters Used AI to Mimic CEO’s Voice in Unusual Cybercrime Case”, Wall Street Journal (30 August 2019) <*https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402*> (accessed 1 February 2021).

(c) 'Bots' being deployed to spread disinformation in advance of elections[6] and in relation to critical public health issues;[7]

(d) German regulators classifying an AI-enabled child's doll as 'illegal espionage apparatus' over fears that it could be used to spy on children;[8]

(e) A number of self-learning 'chatbots', deployed to engage with users on social media, picking up on those users' misogynistic and offensive comments and replying in kind;[9]

(f) Concerns being raised that high-frequency algorithmic trading could facilitate financial market abuse,[10] or – given the interrelatedness of financial markets – amplify systemic risks and precipitate market crashes;[11]

(g) The use of RAI systems to automate highly-targeted cyber-attacks (known as "spear phishing"),[12] thereby substantially increasing the number of individuals or organisations that can be attacked and facilitating "an entirely new scale of cyber-attack";[13] and,

[6] Knight Foundation, *Disinformation, 'Fake News' and Influence Campaigns on Twitter* (October 2018) <*https://s3.amazonaws.com/kf-site-legacy-media/feature_assets/www/misinfo/kf-disinformation-report.0cdbb232.pdf*> (accessed 1 February 2021).

[7] See e.g., Thor Benson, "Twitter Bots Are Spreading Massive Amounts of COVID-19 Misinformation", IEEE Spectrum (29 July 2020). <*https://spectrum.ieee.org/tech-talk/telecom/internet/twitter-bots-are-spreading-massive-amounts-of-covid-19-misinformation*> (accessed 1 February 2021).

[8] Amanda Erickson, "This pretty blond doll could be spying on your family", Washington Post (23 February 2017) <*https://www.washingtonpost.com/news/worldviews/wp/2017/02/23/this-pretty-blond-doll-could-be-spying-on-your-family/*> (accessed 1 February 2021).

[9] Elle Hunt, "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter", The Guardian (24 March 2016) <*https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter*> (accessed 1 February 2021); Chang May Choon, "AI chatbot controversy in S. Korea raises heat about ethics, data collection", Straits Times (24 January 2021) <*https://www.straitstimes.com/asia/east-asia/ai-chatbot-controversy-in-s-korea-raises-heat-about-ethics-data-collection*> (accessed 1 February 2021). These incidents are discussed further at paragraphs 4.32 to 4.34 below.

[10] Mischon de Reya, "Algorithmic trading and market abuse" (26 May 2020) <*https://www.mishcon.com/news/algorithmic-trading-and-market-abuse*> (accessed 1 February 2021).

[11] Technical Committee of the International Organization of Securities Commissions, *Regulatory Issues Raised by the Impact of Technological Changes on Market Integrity and Efficiency – Consultation Report* (July 2011) <*https://www.iosco.org/library/pubdocs/pdf/IOSCOPD354.pdf*> (accessed 1 February 2021).

[12] This is a cyber-attack where an email is tailored to a specific individual or organisation, usually with the intent to steal data or install malware on a target computer or network.

[13] House of Lords Select Committee on Artificial Intelligence, *AI in the UK: ready, willing and able?*, HL Paper 100 (16 April 2018) at [319], [321] <*https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf*> (accessed 1 February 2021).

(h) Concerns regarding the use of "Adversarial AI" – that is, the use of AI-enabled systems to attempt to fool other AI systems into making incorrect classifications or decisions (for example, by making very subtle alterations to pictures, three-dimensional models or signs that the AI system then misrecognises).[14]

1.2 Not all of these examples would or should necessarily incur criminal liability, nor are they raised to suggest that RAI systems are inherently dangerous or detrimental. In many ways, RAI systems will bring significant, sometimes transformative, benefits to society. But where harm to humans, property or wider society does occur, questions necessarily arise as to who is responsible, whether the harm could or should have been prevented, whether those responsible should be sanctioned, and, if so, whether criminal sanctions are appropriate.

1.3 Against that backdrop, this report considers possible risks that the creation and operation of RAI systems pose to humans and property, and examines, when harm does arise, whether and how criminal laws may apply and criminal liability may be attributed[15]. This includes both situations where RAI systems are used to facilitate a human being's wilful criminal acts, and those where it is less straightforward to impute criminal intent to a human actor. The report also studies the extent to which existing laws may or may not be suitable in this regard, and evaluates certain possible alternative or supplementary approaches.

1.4 As RAI systems are still evolving, and are likely to have a near-limitless range of applications, it would not be realistic to attempt an exploration of the contours of criminal liability for all potential uses of RAI systems. Nor, indeed, is a universally applicable, 'one size fits all' approach to the application of criminal liability across all such uses likely to be practicable or appropriate.

1.5 This report therefore does not take a scenario-by-scenario approach, but instead analyses relevant issues through a broad framework focused on two principal vectors: whether or not a human is involved in operating, affecting, or overseeing the RAI system (which generally relates to the degree of automation of that system), and, where such a human is involved,

[14] The concern being that such Adversarial AI could be employed to trick AI-powered cybersecurity systems into allowing malware through firewalls. See *Id.* at [322], [324].

[15] We note that harms resulting from the operation of RAI systems may equally give rise to civil liability (e.g., a private claim for damages). Such issues are beyond the scope of this report. However, certain issues raised in this report – in particular regarding determining to whom liability should be attributed – also arise in a civil liability context. See, in relation to accidents involving autonomous cars, Law Reform Committee, Singapore Academy of Law, *Report on the Attribution of Civil Liability for Accidents Involving Autonomous Cars* (September 2020) (Co-Chairs: Justice Kannan Ramesh and Charles Lim Aeng Cheng) <*https://www.sal.org.sg/Resources-Tools/Law-Reform/Autonomous_Cars*> (accessed 1 February 2021).

whether they intended or knew the harm would occur. Our hope is that that broad framework can be applied to, or can provide guidance regarding, the use of RAI systems across different sectors.

1.6 In light of the above, the remainder of this report is divided as follows:

(a) **Chapter 2** explores the fundamental elements of criminal law and criminal punishment, and how these elements apply to the use of RAI systems.

(b) **Chapter 3** considers the varying degrees of human involvement in RAI-enabled decision-making and the impact of that on applying criminal law.

(c) **Chapter 4** considers the application of criminal liability to users-in-charge of RAI systems and other related parties where operation of those systems results in harm, and the existing and possible alternative or supplementary mechanisms by which criminal law may, where appropriate, be imposed.

CHAPTER 2

THEORETICAL BASES OF CRIMINAL LIABILITY

A ELEMENTS OF CRIMINAL LAW AND CRIMINAL PUNISHMENT

2.1 Criminal law, at its core, concerns the regulation of *human* behaviour. Even where criminal liability is imposed on a corporate entity, regard is ultimately still had to *human* actors,[16] whether as the "directing mind and will" of the company (and thus as embodying the company), or as its agent. With RAI systems expected to replace humans in a range of physical and cognitive roles, there is therefore benefit in examining whether and how the fundamental premises of criminal law continue to be relevant in this new paradigm.

2.2 Broadly stated, criminal law fulfils three aims:[17] to punish the offender; to protect the community against those who cause harm and are dangerous; and to protect offenders by putting them through a system of criminal justice that aims to impose a punishment proportionate to the seriousness of the crime committed. To these ends, criminal law requires three familiar elements that determine whether an accused person should be liable for a crime:

(a) **The accused's conduct, or the "physical element"[18] of an offence** – that is, any fact, proof of which is needed to establish liability under that offence, that is not a fault element of the offence. For present purposes, three aspects of the physical element warrant mention. First, it extends beyond positive actions to omissions or failures to act, where there is a duty to act. Second, it requires the accused's conduct to have been voluntary, and not involuntary. Third, it is not just harm that has been caused that is relevant, but also threatened harm (e.g., liability imposed for dangerous driving, even if nobody has been injured).

(b) **The accused's state of mind, or the "fault element" of an offence.**[19] This refers to any state of mind, proof of which is needed to establish liability under a particular offence,

[16] Particularly in applying the mens rea, or 'fault element', of the offence (see paragraph 2.2(b) below).

[17] David Lanham, Bronwyn Bartal, Robert Evans and David Wood, *Criminal Laws in Australia*, Chapter 1B, "The Purposes of Criminal Law" (The Federation Press, 2006) at 1—4, 7—15 <*http://www.federationpress.com.au/pdf/Lanham%20Ch1B.pdf*> (accessed 1 February 2021).

[18] 22A(2) of the Penal Code (Cap 224, 2008 Rev Ed) ('**PC**'), as introduced by section 6 of the Criminal Law Reform Act 2019 ('**CLRA**')(Act 15 of 2019).

[19] s 22A(1) PC, as introduced by section 6 of the CLRA.

including: (a) intention to cause harm, (b) knowledge that harm is likely to result, (c) wilfulness, (d) rashness in realising that harm may result, and (e) negligence. The fault element (also known as *mens rea*) centres on the accused's mental state, and asks what was in his/her mind at the time of the offence. For present purposes, it should be noted that – in contrast with intention, wilfulness, knowledge, and rashness, which all address the accused's *subjective* state of mind – negligence (i.e., failing to exercise proper care and precaution) is assessed by an *objective* standard.

(c) **Any applicable "exculpatory"[20] defences** that may mitigate or relieve the accused of criminal responsibility. Such defences operate either as "excuses" or "justifications". Excuses (e.g., unsoundness of mind), operate as defences as the law in such a case cannot hold the accused responsible for the harm in question. On the other hand, justifications (e.g., private defence) focus on the nature of the conduct in question.

2.3 In most cases, criminal liability is followed by criminal punishment, which is normally applied pursuant to one or more of the following purposes: (a) retribution, (b) deterrence, (c) incapacitation, and (d) rehabilitation. We do not see a need to elaborate on these purposes – as they are well-understood and are not the primary focus of this report – save to say that it is worth remembering that criminal punishment should be imposed to achieve a judicious and pragmatic balance of these purposes *as a matter of justice and policy* – an objective that can sometimes be forgotten in the heat generated by this topic.

B ISSUES ARISING IN SEEKING TO APPLY CRIMINAL LAW TO RAI SYSTEMS

2.4 Given (a) the focus of existing criminal laws primarily on the regulation of *human* behaviour, (b) the increasing autonomy of RAI systems, and (c) the corresponding reduction in the roles of humans in the decision-making processes of those systems, challenges may arise in attributing criminal liability and holding an individual responsible where harm results from the operation of RAI systems. In particular:

- While criminal liability can be imposed on natural or legal persons – and thus on both humans and corporate entities – an RAI system (such as an intelligent robot) is not a legal person on which criminal responsibility could be placed directly.

[20] Defences may also be 'non-exculpatory', that is, they concede blameworthiness but recognise overriding policy considerations (such as diplomatic immunity). Such defences are outside the scope of this report, however, and not considered further.

- Moreover, every decision made by an RAI system is the result of a long causation chain. The ultimate decision to act in a certain way – which in a non-RAI context would be made by the human undertaking that act – is instead distributed further up the decision-making chain, muddying the identification of blameworthy actors.

2.5 Difficult questions therefore arise regarding not only (a) which aspect of the RAI system factually caused it to act the way it did (resulting in harm), but also (b) which party (or parties)[21] was responsible for that aspect, and (c) whether such party could have foreseen or mitigated the harm.

2.6 Determining those issues may entail a highly complex assessment (factually and technologically), which may not yield a clear outcome. As will be discussed, this may especially be an issue for certain forms of 'machine learning'-based RAI systems, whose algorithms utilise input data to make decisions or predictions, and thus to 'learn' how to complete a task without having to be specifically programmed how to do so.

2.7 More broadly, consideration must be given to whether – from a policy perspective – it is appropriate to apply *criminal* law in a given context or whether, for example, civil liability and/or non-criminal regulatory sanctions may be sufficient or more appropriate. This entails balancing societal imperatives (e.g., encouraging deployers of RAI systems to take necessary safety measures; setting socially-acceptable standards for use of RAI systems; condemning morally unacceptable behaviour) with the desire to avoid disincentivising innovation in RAI system development through the imposition of unreasonable burdens or disproportionate liability exposure.

2.8 Given the wide variety of AI technologies and the myriad applications and settings in which they may be used – each entailing differing sources and levels of possible harm and potential liability (and, indeed, differing levels of potential societal utility and benefit) – a 'one size fits all' approach would not appear practicable. Indeed, the same is true in a non-RAI context, where one finds that different approaches have been taken in different scenarios by policy makers in Singapore. In the context of regulatory rules, for example, examples exist of both:

(a) Criminal sanctions being employed to enforce regulatory compliance, especially where the safety of employees or members of the public is considered by the policy makers to be at risk, or where there is felt to be a need to convey particular moral censure. These range from health products to biomedical research to aviation and construction activities.

[21] For example, the user, the system manufacturer, the system owner, a component manufacturer, a software developer, etc.

(b) Conversely, regulatory compliance being enforced through non-criminal sanctions such as suspension of licences and civil financial penalties (e.g., under the Personal Data Protection Act). The implication here is that non-criminal penalties are viewed as an adequate sanction (in relation to criminal law's moral and punitive function) and as a sufficient deterrent against future misconduct by the sanctioned party and others, without requiring the stigma of a criminal conviction.[22]

2.9 As noted, a similar spectrum is likely to exist in relation to different RAI systems and different scenarios in which they are employed. We would posit that the preferred solution to breaches of the law may well predominantly be a regulatory, non-criminal one, particularly in light of the evolving nature of the technologies involved, and the broader desire to promote innovation in RAI technologies and systems.

2.10 An illustration might be a situation in which an autonomous car breaks the speed limit, or jumps a red light, despite having been designed with the intention that it would comply with road traffic laws,[23] or where RAI-enabled trading systems tacitly engage in collusive practices without the knowledge or intent of their deployers.[24] In such circumstances, a range of non-criminal regulatory sanctions might be imposed, depending on the severity of the breach. These may include improvement notices, civil fines or suspension or withdrawal of approval. Rather than being principally deontological, such regulatory enforcement would serve a more

[22] Thus, for example, in the context of corporate criminal responsibility, the Australian Law Reform Committee has recently opined that civil sanctions on corporations will often be most appropriate, with (broadly stated) criminal liability best limited to scenarios where: (a) it is justified by the level of potential harm that may occur or by broader public interest considerations; (b) civil penalties provide an insufficient deterrent; or (c) denunciation and condemnation of the conduct is warranted. In the ALRC's view: "The criminal law should be applied sparingly so that it retains its core capacity to convey moral opprobrium. The stigma that can attach to the label 'criminal corporation' can be a powerful regulatory tool if the criminal law attaches to serious wrongdoing. Labelling regulatory breaches as 'criminal' where they involve no inherent criminality dilutes the expressive power of the criminal law that makes it such a powerful regulatory tool". Australian Law Reform Commission, *Corporate Criminal Responsibility – Final Report* (ALRC136) (April 2020) at 13 (Recommendation 2) and 1.31. <*https://www.alrc.gov.au/publication/corporate-criminal-responsibility/*> (accessed 1 February 2021).

[23] For example, the car's software may not have been updated to account for recently-changed speed limits or other road rules, or its sensors may have failed to identify a red light in certain novel driving conditions.

[24] The possibility of such 'algorithmic collusion' has been much discussed by competition law practitioners in recent years. See, for example, OECD, "Algorithms & Collusion" <*https://www.oecd.org/competition/algorithms-and-collusion.htm*> (accessed 1 February 2021).

'instrumental' purpose and form part of a feedback loop to 'debug' errors and address unintended 'blind spots'.[25]

2.11 Nevertheless, in other circumstances, the seriousness of the harm or potential harm, the degree of moral culpability or the need for deterrence may justify utilising criminal sanctions or making certain wrongful acts involving RAI systems a criminal offence.

2.12 Before we go on to further consider the basis on which criminal liability may be imposed in such cases, it is useful to briefly explore the extent to which humans and RAI systems may interact in the making of an RAI-enabled decision. This, and its implications for the imposition of criminal liability, will be explored in the next chapter.

[25] Simon Chesterman, "Artificial Intelligence and the Problem of Autonomy" (2020) 1(2) Notre Dame Journal on Emerging Technologies 210 at 226.

CHAPTER 3

THE EXTENT OF HUMAN INVOLVEMENT IN RAI-ENABLED DECISION-MAKING AND THE EFFECT ON CRIMINAL LAW

A DECISION-MAKING MODELS

3.1 The Personal Data Protection Commission's *Model AI Governance Framework*[26] posits three broad approaches on the extent of human oversight over RAI-enabled decisions. These are:

(a) **Human-in-the-loop ("HITL")**. The HITL approach involves a human decision-maker who relies on the intelligent system to suggest one or more possible options. The human, however, makes the final decision to proceed with the relevant action. As humans still remain an integral part of the decision-making process, they are "in the loop".

(b) **Human-over-the-loop ("HOVTL")**. In a HOVTL system, the human is not directly involved in deciding or carrying out the relevant decision, but retains oversight over the entire decision-making process. He or she might even influence the process by adjusting the parameters of decision-making during the operation of the algorithm.

The operation of a GPS system is one such example: the system decides between possible routes to get from one point to another, but the driver retains oversight (in ultimately deciding whether to follow that route) and influence (in altering the parameters, such as for unforeseen road obstructions).

(c) **Human-out-of-the-loop ("HOOTL")**. Under the HOOTL approach, within set parameters, the AI system is responsible for all aspects of the making and execution of the decision, without any human input or interaction.

An example is the decision-making of an autonomous cleaning bot, which maps out and executes the best path for cleaning a location, excluding "no-go zones" pre-determined by humans.

3.2 As can be seen, the choice of decision-making model determines the extent of oversight, control and responsibility the human user has over/for the decision in question.

[26] Personal Data Protection Commission of Singapore, *Model AI Governance Framework (Second Edition)* (21 January 2020) at [3.14], <*https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf*> (accessed 1 February 2021).

B LEVELS OF AUTOMATION

3.3 A similar schematic can be seen in the various classifications that have been adopted to define an RAI system's level of automation, and which describe the shifting roles of humans and RAI systems as such automation increases.

3.4 While these models differ in their specific classification of different levels of automation, and in the specific nomenclature adopted (which is to be expected given that they refer to different industries involving different actions and skillsets), they each describe a graduated transition from manual to automated operation. In this way, they help provide, to some degree, a 'common language' on which laws and policies can start to be built. Below, we set out various examples of these models.

3.5 The first model – now widely adopted by regulators and stakeholders in the transport sector[27] – is the scale of automation developed for automated vehicles by the Society of Automotive Engineers International ('**SAE**').[28] This six-level classification, summarised in the table below, ranges from "no automation" (Level 0), where the human driver performs all driving functions, to "full automation" (Level 5), where the vehicle can drive itself anywhere a human driver can, with no human input.

Level	**Description**
0	**No automation**. A human driver performs all aspects of all driving tasks, even though these can be enhanced by warning or intervention systems.
1	**Driver assistance**. The driver assistance feature(s) can carry out either steering or acceleration and deceleration.
2	**Partial automation**. Various driver assistance features can combine to carry out both steering and acceleration or deceleration. The driver is responsible for monitoring the driving environment and must remain actively engaged in the driving task.
3	**Conditional automation**. The driving automation features are generally capable of performing all driving tasks but the human driver, as a "fallback-ready user", is expected to respond appropriately to any request to intervene. Thus, while the driver is not expected to monitor the driving

[27] See further *Report on the Attribution of Civil Liability for Accidents Involving Autonomous Cars*, above, n 15 at [1.16].

[28] Society of Automotive Engineers, *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, SAE International J3016_201806 (June 2018) <*https://www.sae.org/standards/content/j3016_201806/*> (accessed 1 February 2021).

	environment, he must be receptive and responsive to a handover request or to an evident system failure.
4	**High automation**. The driving automation features can perform all the driving tasks even if a human driver does not respond to a request to intervene. If the limits of the autonomous driving system are, for whatever reason, exceeded, the system will respond by putting the vehicle in a "minimal risk condition" (e.g., by coming to a gradual stop, or changing lanes to rest on the road shoulder).
5	**Full automation**. The vehicle is capable of performing all driving functions in all situations and conditions that a human driver can.

3.6 The second model is Sheridan and Verplank's ten-level taxonomy developed in 1980, which focused on human-computer decision-making:[29]

Level	**Description**
1	Human considers alternatives, makes and implements decision.
2	Computer offers a set of alternatives which human may ignore in making decisions.
3	Computer offers a restricted set of alternatives, and human decides which to implement.
4	Computer offers a restricted set of alternatives and suggests one, but human still makes and implements final decision.
5	Computer offers a restricted set of alternatives and suggests one, which it will implement if human approves.
6	Computer makes decision but gives human the decision to veto prior to implementation.
7	Computer makes and implements decision but must inform human after the fact.
8	Computer makes and implements decision and informs human only if asked to.

[29] Thomas Sheridan, "Computer Control and Human Alienation" (1980) 81 (1) Technology Review 60, as cited in Jorgen Frohm, Veronica Lindstrom, Mats Winroth & Johan Stahre, "Levels of Automation in Manufacturing" (2008) 30(3) International Journal of Ergonomics and Human Factors 19.

9	Computer makes and implements decision and informs human only if it feels this is warranted.
10	Computer makes and implements decision if it feels it should and, if so, only informs human if it feels this is warranted.

3.7 The third and final model we reference is Frohm et al's 'Levels of Automation in Manufacturing'.[30] This model consists of seven levels, and progresses from a task performed entirely manually to one performed fully automatically:

Level	**Description**
1	Totally manual: Totally manual work done by the worker's own muscle power in which no tools are used.
2	Static hand tool: Manual work with the support of a static tool (e.g., screwdriver).
3	Flexible hand tool: Manual work with the support of a flexible tool (e.g., adjustable spanner).
4	Automated hand tool: Manual work with the support of an automated tool (e.g., hydraulic bolt driver).
5	Static machine / workstation: Automatic work by a machine that is designed for a specific task.
6	Flexible machine / workstation: Automatic work by a machine that can be reconfigured for different tasks.
7	Totally automatic: Totally automatic work in which the machine solves all deviations or problems that occur by itself (e.g., autonomous systems).

3.8 Two brief observations may be made. First, while each of the three models categorise and delineate the extent of automation somewhat differently according to the relevant industries' needs and circumstances, each can nonetheless be condensed, in effect, into a three-level spectrum of automation:

[30] Jorgen Frohm, Veronica Lindstrom, Mats Winroth & Johan Stahre, "Levels of Automation in Manufacturing" (2008) 30(3) International Journal of Ergonomics and Human Factors 19.

(a) manual;

(b) partial automation (automation with human involvement); and

(c) full automation (with no human involvement).

While it is not intended to carry legal weight, we believe that this spectrum provides a useful tool for analysis of the extent of human involvement at a given point in a specific RAI system's operation, and – in turn – in assessing the level of responsibility (if any) of that human user for that system's actions. We therefore use this spectrum as the basis of our analysis in Chapter 4 below.

3.9 Second, it can be seen that this automation spectrum broadly correlates with the decision-making model articulated by the PDPC. HITL and HOVTL decision-making systems, in which humans remain a part of the overall decision-making process, can be classified as 'partial automation'. By contrast, HOOTL systems, in which humans are removed from the overall decision-making process, would represent 'full automation'.

3.10 A further example of an equivalent dichotomy can be seen in the Law Commission and Law Commission of Scotland's (together, the "**UK Commission**") work on automated vehicles.[31] The UK Commission distinguished between circumstances in which the so-called "user-in-charge" was driving the car, and those where the automated driving system ('**ADS**') was engaged. It will be noted that those two states of operation – one where the user is controlling the vehicle, and another where he or she is not – broadly correspond with the two key aspects of the automation spectrum highlighted above (i.e., partial automation with some degree of human involvement, and full automation). What is particularly notable in the UK Commission's report for our purposes is the impact of those differing states on criminal liability. In the former, the user-in-charge would be liable in criminal law for any infringement of road rules or standards; in the latter, criminal liabilities on that user-in-charge would (broadly stated) be replaced by regulatory or legal liabilities placed on the entity(-ies) responsible for the automated system (the '**ADSE**').[32, 33]

[31] Law Commission of England and Wales, "Automated Vehicles" <*https://www.lawcom.gov.uk/project/automated-vehicles/*> (accessed 1 February 2021).

[32] UK Commission, *Automated Vehicles: Consultation Paper 3 - A regulatory framework for automated vehicles* (Law Commission Consultation Paper 253)(18 December 2020) ('***UK AV Consultation***') <*https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2021/01/AV-CP3.pdf*> (accessed 1 February 2021). Specifically, the UK Commission proposes (at e.g., [8.49] – [8.50]) that "where the ADS acts in a way which (if done by a human driver) would lead to criminal or civil liabilities, the ADSE would be subject to regulatory action under the safety assurance scheme. The aim would be to stop mistakes from happening again. There would therefore be gradated sanctions, including improvement notices, fines and (in the most serious cases) recalls." As can be seen, the UK Commission's consultation also addresses the attribution of civil liability. Such issues are beyond the scope of this report, but are

(*cont'd on the next page*)

3.11 As will be discussed further below, we consider that such a transition from 'user responsibility' to 'automated system entity responsibility' as the degree of automation increases may be applicable to RAI systems more broadly.

discussed in another report in this series: *Report on the Attribution of Civil Liability for Accidents Involving Autonomous Cars*, above, n 15.

[33] With reference to the discussion in paragraphs 2.7 – 2.11 above, it is notable that the UK Commission considered it generally more appropriate to impose non-criminal regulatory sanctions on ADSEs. While it has proposed the creation of new criminal offences applicable to ADSEs, these are limited to situations of "serious wrongdoing" in which ADSEs dishonestly omit safety-relevant information or include misleading information when putting a vehicle forward for classification as self-driving or responding to information requests from the regulator. *UK AV Consultation*, *id.* at [14.13] – [14.21] and [14.109]. See further paragraph 4.61 below.

CHAPTER 4

LIABILITY OF USERS-IN-CHARGE AND OTHER PARTIES

4.1 Against the backdrop described in the previous chapters, it will be seen that, whether – and to which legal person – criminal liability for a harmful act involving an RAI system should be attributed is, broadly stated likely to be a function of:

(a) the severity and risk of actual or potential harm inherent in the use of the system in the relevant context;

(b) the level of automation of that system; and,

(c) the degree of human oversight over, and involvement in, that system's decision-making.

4.2 Of particular interest in determining who should be liable is the interaction of (b) and (c). As noted above, two discrete scenarios can be differentiated.

- One, where the level of automation of the RAI system in question is more limited and/or the degree of human oversight greater (that is, "partial automation"). Thus, there is a need to examine the potential liability of the human 'user-in-charge'.
- The second, involves so-called "full automation" or "human-out-of-the-loop systems" (i.e., those which involve no human input or interaction in the RAI system's making and execution of a decision within set parameters), where there may be no such identifiable human user involved. Thus, the question arises as to which other person or entities (if any) might be criminally liable, and on what basis.

We consider each of these scenarios in turn below.

4.3 Under the first of the above scenarios, the first inquiry is to identify what we hereafter refer to as the 'user in charge'[34] – that is, the human involved in the operation of the RAI systems who:

- for RAI systems with only a limited degree of automation, [35] directly controls or is responsible for determining or the actions of the RAI system; or,

[34] While utilising the same term, the definition of 'user-in-charge' adopted here differs from that utilised by the UK Commission in the specific context of automated vehicles (although its 'users-in-charge' would equally fall within the definition utilised here).

[35] For example, AI-enhanced tools that merely assist the human user in the relevant task.

- for more highly-automated (but still not fully automated) systems:
  - continues to bear ultimate responsibility for deciding on or approving a particular action; or,
  - even though not responsible for taking the decision or for carrying it out, nonetheless (i) retains oversight over the RAI system's entire decision-making process or (ii) is under a specific duty to intervene to control the RAI's actions in given scenarios.

4.4 In many circumstances – particularly at lower levels of automation – the identity of the user-in-charge in any given instance will be self-evident. In others (for example, in relation to more highly-automated vehicles), it may be necessary or appropriate to define expressly in laws or regulations:

(a) who that user-in-charge is;

(b) the extent of responsibilities on them to oversee or control the RAI system's behaviours;[36] and/or,

(c) if necessary, a specific point beyond which the RAI system is considered to be operating autonomously, such that the human is no longer considered 'in-charge' and thus would not be considered liable for offences that occur.

4.5 Where such a human 'user-in-charge' remains involved in the use or operation of an RAI system, a further distinction can be drawn for present purposes between:

(a) Cases of *intentional* criminal use of, or interference with, the RAI system; and,

(b) Cases where *non-intentional* criminal harm is caused.

This dichotomy is common in criminal law. Most criminal laws will focus on the former and require some degree of intentional, knowing or wilful conduct. Nonetheless, there remain various examples where criminal liability may be found even where such intention was lacking, primarily offences where rashness or negligence is sufficient, and strict liability offences (where the state of mind is irrelevant).

[36] Such responsibilities would necessarily be calibrated to what it would be reasonable or practicable for a human user-in-charge to do in any given scenario to, for example, assess the risks and/or take remedial action. This issue, as it applies to autonomous cars, is discussed in detail in the *UK AV Consultation*, above, n 32 (at, e.g., Chapters 3 and 4).

A INTENTIONAL CRIMINAL USE OF OR INTERFERENCE WITH RAI SYSTEMS

4.6 As regards cases of intentional criminal use of, or interference with, the RAI system, it appears uncontroversial as a general policy principle that where a person (including a third party who is not the user-in-charge) *intentionally* uses (or causes) an RAI system to cause harm, that person should be liable for the act.

4.7 In our view, there will be many circumstances in which existing laws could be applied to sanction incidents of such intentional misuse or interference.

4.8 For example, the Computer Misuse Act (‘**CMA**’) (which is focused on intentional criminal misuse of computers) provides, inter alia, that:[37]

> ***Access with intent to commit or facilitate commission of offence***
>
> **4.**—(1) Any person who causes a computer to perform any function for the purpose of securing access to any program or data held in any computer with intent to commit an offence to which this section applies shall be guilty of an offence.

4.9 This offence would likely catch conduct such as a person hacking into an RAI system to modify its algorithms or models, so as to cause the RAI system to cause damage to a person or property.

4.10 Similarly, recourse might be had to the Penal Code (Cap 224, 2008 Rev Ed) (‘**PC**’). More specifically, for example, sections 324 and 326 PC prohibit voluntarily causing (grievous) hurt by dangerous weapons or means. Such prohibitions would appear on their face capable of covering most situations of an offender voluntarily programming or using an RAI system as a weapon to inflict physical hurt on a person. This is evident from the broad phrasing of terms such as “by means of any instrument for shooting, stabbing or cutting or any instrument which, used as a weapon of offence, is likely to cause death”[38] which appear sufficiently wide to capture RAI systems. Potential examples of such conduct might include a person programming an automated drone to shoot a projectile at a victim or altering a robotic pet to injure someone.

4.11 However, it is not clear that the CMA, Penal Code or other legislation would fully or adequately cover certain other situations of intentional malicious use of or interference with RAI systems (or, at a minimum, certain provisions or definitions in those statutes may become stretched, or not sit easily, when extended to such RAI systems). By way of example:

- While the use of RAI systems to conduct “spear phishing” attacks or other conventional cyberattacks (e.g., denial-of-

[37] Computer Misuse Act (Cap 50A, 2007 Rev Ed), s 4(1).

[38] S 324 and 326 PC.

service or malware attacks) would be covered under the various provisions of the CMA, it is possible that new forms of cyberattacks committed through RAI systems may not be caught by existing laws. It is less clear, for instance, how adversarial AI attacks would be covered by the CMA.

- Questions may also arise, for example, as to whether the criminal law adequately catches the nexus or "causal link" between the human perpetrator and the RAI system employed to cause that harm, or whether new, emergent technologies fall within ss 324 and 326 PC's existing notions of "instrument" or "weapon of offence".
- Similarly, consider a situation where a person intentionally obstructs the signals to an RAI system's sensors (or sends misleading signals), such that it fails to 'see' and collides with a person or other object.[39]
    - Certainly, it would seem that section 4(1) of the CMA does not extend (and was not envisaged to extend) to such conduct.
    - It is arguable that such interference – where intended or known to be likely to cause injury, fear or annoyance – could amount to the use of criminal force, contrary to section 350 PC,[40] even if the harm did not eventually materialise or a different harm was eventuated.
    - But here too, the limitations of the law quickly reveal themselves. Prime among these is the fact that section 350 PC only covers acts resulting in injury (physical), fear or annoyance to the victim, whereas the use of RAI systems could equally result in damage other than physical harm, such as financial loss.[41]
    - While in certain contexts such types of harm may be covered by specific pieces of legislation (e.g., where the interference with the RAI system is used to manipulate financial markets), one could equally envisage harm beyond those specified in section 350 arising in an unregulated domain.

[39] *UK AV Consultation*, above, n 32 at [15.1] – [15.11].

[40] "Whoever intentionally uses force to any person, without that person's consent, in order to cause the committing of any offence, or intending by the use of such force illegally to cause, or knowing it to be likely that by the use of such force he will illegally cause injury, fear or annoyance to the person to whom the force is used, is said to use criminal force to that other."

[41] Furthermore, for example, insofar as section 350 is a general provision, the punishment prescribed in s 352 PC (imprisonment for up to 3 months and/or a fine not exceeding $1,500) may not be proportionate to or adequate for the range of conduct that could be intentionally caused through interference with RAI systems.

4.12 In light of the above, we consider firstly that there is merit in reviewing existing laws with a risk-based lens to identify and fill gaps that may exist in relation to RAI systems, to ensure that intentional harms caused through the use of such systems are caught.[42]

4.13 Secondly, insofar as many existing regulations in effect protect against the *effects* of RAI abuse, not the *cause* itself, we note also that in particular areas there may remain a specific need for, or benefit in, new legislation to introduce criminal offences crafted to deal with certain intentional harms inflicted through the use of RAI systems 'at source'. Needless to say, as with any new form of criminal liability, the scope of the harm covered by such an offence would require clear definition (so as not to be left unduly open-ended) and careful calibration.

4.14 A recent example of this[43] (which usefully demonstrates the Singapore Parliament's responsiveness in adapting its laws to address emergent risks and the potential challenges of formulating such targeted legislation) relates to the increasing use of AI-enabled "bots".[44]

4.15 Specifically, concerns have arisen regarding the ability of such bots to accelerate the creation and dissemination of disinformation (colloquially-termed "fake news") and to incite unrest or hatred, including by simulating user behaviour on social media platforms and responding to other users' postings based on scripts that they have been programmed to use.[45] In response to such concerns, Parliament introduced in section 8(1)

[42] By way of example, the Singapore Government recently proposed updates to firearms legislation to address the threat of, among other things, armed automated drones. In particular, the new laws would clarify that a person who "drives, flies or otherwise operates (even by remote control) any vehicle, vessel, aircraft or other device conveying or otherwise carrying" a weapon is treated as possessing them (Guns, Explosives and Weapons Control Bill (Bill 44 of 2020), cl 5(i)).

[43] Note also the Penal Code Review Committee's ('PCRC') consideration of a new criminal offence of making, altering, or using a computer program so rashly or negligently as to be likely to cause harm (discussed further at paragraph 4.48 to 4.54 below). The PCRC noted expressly that the offence would be targeted at risk-creation and thus could be committed regardless of whether any hurt or injury was in fact caused. PCRC, *Report of the Penal Code Review Committee* (August 2018) ('***PCRC Report***') at 30 <*https://www.mha.gov.sg/docs/default-source/default-document-library/penal-code-review-committee-report3d9709ea6f13421b92d3ef8af69a4ad0.pdf*> (accessed 1 February 2021).

[44] S 2(1) of the Protection from Online Falsehoods and Manipulation Act 2019 (Act 18 of 2019). defines a "bot" to mean a computer program made or altered for the purpose of running automated tasks.

[45] Center for Information Technology and Society at UC Santa Barbara, "How is Fake News Spread? Bots, People like You, Trolls and Microtagging" (29 Aug 2018) <*https://www.cits.ucsb.edu/fake-news/spread*> (accessed 1 February 2021). In one high profile example in Canada, a bot designed to sow mistrust achieved its purpose by spreading false claims on vaccination, seeding doubts regarding vaccination that it is believed contributed (at least in part) to the recent increase in cases of measles globally. See, Andy Blatchford, "False Vaccine Claims Spread Online by Bots, Trolls:

(*cont'd on the next page*)

of the Protection from Online Falsehoods and Manipulation Act ('**POFMA**'),[46] which came into force in October 2019, a specific offence of "mak[ing] or alter[ing] a bot with the intention of:

(a) communicating, by means of the bot, a false statement of fact in Singapore; or.

(b) enabling any other person to communicate, by means of the bot, a false statement of fact in Singapore."[47]

4.16 Section 8(1) is notable, insofar as it targets the *production, manufacture or alteration* of a "bot" rather than the communication of the false statement itself. It is therefore targeted at the intentional human actions applied to AI technology that led to the harmful act (thus, seeking to tackle the harms 'at source'). Also notable is the calibration of punishments for the offence with the extent of (potential) harm caused: pursuant to section 8(3) POFMA, higher penalties can be imposed if the communication of the false statement has certain prejudicial effects such as influencing the outcome of a political election, or inciting feelings of enmity, hatred or ill will.

4.17 Several points can be made. First, in focusing on the production of the bot, rather than the communication, the Act implicitly recognises that the bot is itself capable of autonomously issuing a false statement without any human intervention. Yet in doing so, it does not address how such autonomously-made false statements should be addressed if the creator of the bot did not specifically intend for them to be made.[48] Second, at the time of writing, we are not aware of the Act having been invoked against misinformation spread by "bots" and other forms of RAI technology. Third, the legislative objective of the Act is to combat false information and manipulation online. To the extent that bots engage in other forms of harmful conduct outside the Act's intended scope (such as engaging in hate speech or cyberbullying/harassment), such harms will need to be addressed under other existing or newly-created laws. Last, we note that the Act's broad, technology-neutral definition of 'bots'[49] should sufficiently "future-proof" the legislation to cover RAI technology developments in the foreseeable future. Nonetheless, given the rate of technological change, that definition (and equivalent definitions of RAI systems in other new

Top Public-health Doc", CBC (26 April 2019) <*https://www.cbc.ca/news/elections/false-vaccine-spread-by-bots-trolls-1.5113716*> (accessed 1 February 2021).

[46] Above, n 44.

[47] POFMA, above, n 44, s 8(1).

[48] As was the case in relation to Microsoft's deployment of its 'Tay' chatbot and the 'Luda Lee' chatbot recently launched on Facebook by the South Korean company Scatter Lab (each discussed further at paragraphs 4.32 to 4.34 below). While those bots (contrary to their designers' intentions) engaged in offensive and hate speech, it is not difficult to imagine that they could equally have 'learnt', based on their engagement with other users, to make false statements of fact or amplify misinformation.

[49] See n 44 above.

legislation) will necessarily have to be monitored and fine-tuned over time as unanticipated new technology emerges.

B NON-INTENTIONAL CRIMINAL OFFENCES AND RAI SYSTEMS

4.18 The challenges regarding the attribution of criminal liability become rather more pronounced where harms result from the RAI system's operation, even though it cannot be said that the user-in-charge (if there is one) or any other (natural or legal) person *intended or knew* that such harm would occur. Is it appropriate to impose criminal liability in those circumstances (which, hereafter we term 'non-intentional harms')? And if so, on whom?

4.19 As regards the appropriateness of criminal liability, this will – as noted at paragraph 2.7 above – ultimately be a question of policy, taking into account various moral, social and other imperatives, and is likely to vary depending on the degree of actual or potential harm caused, and the circumstances in which the RAI system was deployed. In many cases, civil or regulatory enforcement may be considered more appropriate.

4.20 Nonetheless, even under existing criminal laws (notwithstanding that they are still generally rooted in a 'fault-based' framework), it is possible for the fault element of certain offences to be satisfied – and thus criminal liability imposed – even where a (natural or legal) person did not intend the harm.

- Prime among these are various offences under the Penal Code that may be committed where that person was negligent, as well as where the offence was committed intentionally, knowingly or rashly.[50]
- In the context of the discussion in the preceding paragraph, both the general[51] and context-specific[52] negligence-based offences in the Penal Code generally require acts that cause *serious* harm or are inherently dangerous to human life or personal safety, and have potential punishments that are relatively low.[53]

4.21 In this chapter, we therefore consider first how such criminal negligence principles may apply in the context of non-intentional harms resulting from the operation of RAI systems, including the challenges that arise. In the light of those challenges, we then evaluate certain possible

[50] Thus, for example, s 304A and s 338 PC impose criminal liability on a person who causes death (s 304A) or grievous hurt (s 338) by their rash or negligent act.

[51] See s 304A(b), 338(b), 337(b) and 336(b) PC.

[52] See s 269, 279 – 280, 282 and 284 – 289 PC.

[53] *PCRC Report*, above, n 43 at 178.

alternative (or supplementary) means to define and attribute criminal liability.

4.22 The intention here is not to specifically advocate for one solution over another. Indeed, as noted above, no single approach is likely to be appropriate to cover the full range of RAI system applications and potential harms. Rather, we have sought to highlight the merits and challenges inherent in those approaches, and – as appropriate – the scenarios in which they might be more (or less) effective.

1 Addressing non-intentional harms through criminal negligence

4.23 Certain harms resulting from the operation of RAI systems may fall within the scope of the existing general or context-specific negligence offences presently in the Penal Code. Thus, for example, section 287 PC imposes criminal liability on a person who negligently uses, or fails to take due care with, machinery in his possession or under his care, in such a way as to endanger human life, or to be likely to cause hurt or injury to a person. Many types of deployed RAI system are likely to be deemed "machinery" for these purposes, and thus it may be possible to sanction the negligent human user of an RAI system under such existing laws.

4.24 However, certain limitations and challenges of applying provisions such as section 287 PC to RAI systems soon become apparent. Not least:

(a) many other forms of RAI system – such as AI software – may not necessarily fit easily within the scope of the term 'machinery' for the purposes of section 287; and

(b) as noted above, those existing offences typically focus on acts that are inherently dangerous or pose a danger to human life or personal safety. However, in the context of RAI systems, there may be circumstances in which negligent behaviour results in severe harm, and therefore criminal liability might be deemed appropriate, even though human life or personal safety is not directly endangered. Examples might include widespread and critical disruptions to broadband network services, or severe systemic risks to banking and financial trading systems. In its 2018 Report, the Penal Code Review Committee alluded to this issue, noting that "computer programs" (which would include algorithms and AI software) "are able to cause types of harm that machinery cannot."[54]

4.25 However, seeking to apply the Penal Code's criminal negligence provisions to harms resulting from the operation of RAI systems also exposes the more fundamental challenges of applying such a broad, fault-

[54] *Id.* at 29.

based framework in novel contexts. 'Negligently', for the purposes of the Penal Code, is defined as follows:

> *Whoever omits to do an act which a reasonable person would do, or does any act which a reasonable person would not do, is said to do so negligently.*[55]

Inherent in that definition is the need to both (a) determine what an objective 'reasonable person' would do in a given circumstance, and (b) prove (beyond reasonable doubt) that that standard was breached in the case at hand – in essence that the person in question was 'at fault'.

4.26 Where a harm resulting from the operation of an RAI system falls within the scope of the existing negligence-based offences in the Penal Code, it would be for the courts in individual cases to determine what a reasonable person would or would not do. In so doing, the courts would apply or adapt existing criminal negligence standards, or – in the absence of precedent – define new ones. The same would likely be true of any other, newly-created and generally applicable (as opposed to application- or sector-specific) negligence-based offence introduced specifically to cover harms from RAI systems. This has the benefit of enabling the law to adapt to a range of circumstances. However, it is also inherently uncertain, in particular as RAI systems may give rise to forms of conduct for which existing precedents are inappropriate, or for which there is no existing precedent at all. The concern is that such uncertainty could have an unintended chilling effect on innovation in, and the development or adoption of, potentially beneficial RAI systems.

4.27 Given those risks, an alternative would be to set out the nature and extent of the relevant standard of conduct more precisely in sector- or technology-specific legislation, rather than leaving it to the courts to establish them over time. For example, in an autonomous vehicle context, legislation might impose a requirement on a user-in-charge of a vehicle to take over control of the vehicle in defined circumstances (e.g., where a police officer directs the vehicle to stop or where a road is temporarily closed by a traffic accident, etc.). Or that legislation might specify required actions regarding the need to maintain the vehicle's sensors in good working condition or update its software or data sets. Such an approach would help address the uncertainty inherent in more general negligence offences. However, the flip side of this is that it is not practicable to legislate such a standard for all possible applications of RAI systems, or all possible risks of harm that may arise. This is especially so in sectors where RAI technologies and their applications are evolving rapidly.

4.28 Establishing the nature and extent of the standard of conduct that exists (i.e. what a reasonable person would/would not do) is not the only

[55] S 26F(1) PC. This provision codified the definition of "negligence" in *criminal* contexts previously articulated by the High Court in *PP v Hue An Li* [2014] 4 SLR 661 at [38].

challenge, however. Even where that can be established, it must then be shown that that standard has been breached.

4.29 Where the standard is imposed on a human, for example as a user-in-charge of an RAI system, that may be relatively straightforward to demonstrate. In particular, the forensic investigation into the human's actions would not be fundamentally different from any other criminal negligence investigation.

4.30 The greater challenge arises where the human user has not been negligent, and yet some harm is still caused by the way the RAI system operates. Or, for equivalent reasons, where such harm results and there is no human user at all.

4.31 There may be various reasons, unrelated to the conduct of the user, why such harm was caused or the RAI system otherwise failed to comply with certain criminal laws. Sometimes, those reasons may be clear, and readily addressable by the law. The harm could, for example, be the result of actions taken by a person other than the user-in-charge, such as a person making unauthorised modifications to an RAI system without the user-in-charge's knowledge. In those circumstances, liability is most appropriately placed on the person having taken those actions, based on existing criminal law principles – including, where appropriate, criminal negligence principles.[56]

4.32 However, in other circumstances, the reason for the RAI system's actions, and thus what caused the harm, may not be so apparent. For example, the cause of the harm might lie in the RAI's software itself. In that case, greater complexities are liable to arise.

- As noted above, the process by which an RAI system's software determines how the system should act in any given scenario is highly complex (especially in the case of 'deep learning'[57] systems), and may be difficult to establish definitively. Every stage of the AI deployment process – the data preparation, the training of the model, the choosing of the relevant model(s), the environment in which it is deployed and so on – could have played a role in the decision eventually made by the RAI system.

[56] E.g., that person would need to be shown to have had the necessary *mens rea* for the offence in question, whether that be intention, knowledge, negligence, etc. The analysis here is the same as discussed above in relation to intentional harms.

[57] Deep learning is a specific form of machine learning that utilises 'artificial neural networks' to model and draw insights from complex structures and relationships between data and datasets. Briefly stated, artificial neural networks are a series of 'layered' algorithms used to analyse, classify, learn from and interpret input data. The values from one layer are fed into the next layer to derive increasingly refined insights.

- For example, it might depend not only on the RAI system's underlying code, but also on the quantity, quality and accuracy of the data on which the system's learning was based, the comparability of the environment in which it was trained with that in which it is ultimately employed, or the particular real-world data it received at a given point in time.
- Even if that can be established, questions then arise as to who was responsible for that aspect of the RAI system's decision making, and whether they acted in a way which would justify the imposition of criminal liability (whether under a criminal negligence or other standard).
- These questions are liable to be even more difficult to resolve where the RAI systems utilise deep learning to 'adapt' and fine-tune their decision-making processes.[58] A renowned example of such learning and adaptation resulting in detrimental outcomes is that of the "Tay" chatbot created by Microsoft to respond to social media users' queries and become progressively 'smarter' in its responses. After Tay went live on Twitter and became exposed to human users in a real-life environment, it 'went rogue' – picking up on users' misogynistic and offensive comments, and replying in kind.[59] Similar problems have since arisen after the deployment of other self-learning chatbots. Earlier this year, for example, within weeks of being deployed on Facebook to interact with South Korean users, the "Luda Lee" chatbot began making offensive comments about disability and homosexuality, and sharing people's personal information.[60] Due to the nature of their 'deep learning', the precise way those chatbots altered their output through exposure to real-life conversation could not be specifically predicted (or, the makers of Luda Lee

[58] For example, consider the scenario in which a deep-learning RAI system, as well as being 'trained' by the developer during testing, continues to fine-tune its decision making after deployment, while under the 'control' of the user. In those circumstances, was the particular harmful action that the system took (which, we note, will in fact likely be the combined result of myriad individual 'decisions' made by the system) a result of something it 'learned' before its deployment, or after? Can that be established? And what effect should it have on which person(s) or entity(ies) are deemed 'responsible' for that action?

[59] Elle Hunt, "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter", The Guardian (24 March 2016) <*https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter*> (accessed 1 February 2021).

[60] Chang May Choon, "AI chatbot controversy in S. Korea raises heat about ethics, data collection", Straits Times (24 January 2021) <*https://www.straitstimes.com/asia/east-asia/ai-chatbot-controversy-in-s-korea-raises-heat-about-ethics-data-collection*> (accessed 1 February 2021). Further recent examples include a Japanese chatbot that expressed Nazi sympathies, and 'Simsimi', another chatbot launched in Korea, that swore at users.

claimed, prevented) in advance. Nor could it necessarily be fully explained (e.g., as a sequence of logical 'decisions') after the event.

4.33 This latter issue is commonly known as the "black box" problem, where it can be challenging (if not impossible) for humans to comprehend the decisions that RAI systems reach, in particular through continuous training.[61] Beyond the practical challenges of determining 'causes' of harm, this indeterminable element of RAI systems also creates difficulties for the assessment of liability. It may not be possible to predict with a reasonable level of certainty how an RAI system would act when developed or used in a certain way, making it difficult, in turn, to discern "what a reasonable person should (or should not) do" to prevent the harm from arising.

4.34 For example, let us assume hypothetically that Tay's operation had caused harm for which there was potential liability on criminal negligence grounds. What standard of care would be imposed on the person deploying Tay to prevent that harm from arising, below which criminal negligence liability might be imposed? Short of not deploying the system at all, how would a reasonable deployer react to a continuously-evolving RAI system? Would they be expected to maintain continuous oversight and verification of the system so that it could be shut down when it started to 'malfunction'? If so, does that risk negating any benefit the deployer might have obtained from deploying the chatbot?

4.35 Setting aside broader questions of whether, in a particular circumstance, it is appropriate to impose criminal liability at all, the preceding discussion shows some of the challenges of applying a fault-based negligence framework to try to assess and attribute such criminal liability:

(a) where the user in charge cannot be said to have been at fault; or,

(b) in 'human-out-of-the-loop' scenarios, where there is no direct human involvement in or oversight of the RAI system's operation at all (and thus where the question of a human user's level of culpability does not arise).

4.36 Ultimately, the challenge with reliance on criminal negligence lies in its uncertainty. In particular, given:

(a) the multiple players involved in developing, deploying and using AI;

[61] See, Ronald Yu and Gabriele Spina Ali, "What's Inside the Black Box? AI Challenges for Lawyers and Researchers" (2019) 19(1) Legal Information Management 2, at 5; Dave Gershgorn, "AI is now so complex its creators can't trust why it makes decisions", Quartz (7 December 2016) <*https://qz.com/1146753/ai-is-now-so-complex-its-creators-cant-trust-why-it-makes-decisions/*> (accessed 1 February 2021).

(b) the different forms of RAI system; and,

(c) the potential ability of the RAI system to learn from its surroundings and produce unexpected and even unexplained outcomes,

it may be a challenge to determine not only which entity was responsible for the system's actions, but – even if that could be established – what would count as an entity falling short of reasonable standards, and thus as criminal negligence. If expectations of what users or entities should do to prevent or address harm are set too low, the law loses any real force. However, if that bar is set too high, there is a risk that potentially beneficial deployment of RAI systems will be deterred.[62]

4.37 There is therefore merit in considering whether other mechanisms for applying criminal law may be preferable to, or could usefully supplement, reliance on criminal negligence.

4.38 Before doing so, however, we note as an aside that the challenges of understanding how a RAI system arrived at its decision – through the data used, processes undertaken or algorithm chosen – have increasingly led to calls for greater emphasis on, and research into, "explainable AI".[63,64] Briefly, this is a growing area of research into the development of tools and processes to explain how RAI systems function and arrive at decisions and/or predictions. While acknowledging that explainability is not, in and of itself, a panacea, we support those calls:

- Such explainable AI tools could provide greater clarity as to the reasons for which an RAI system made a particular decision (such as due to certain pieces of data being used to train it).
- Equivalent tools to enhance 'traceability' (that is, ensuring an RAI system's decisions and the datasets and processes that yield them are documented) might similarly provide further insight into, for example, human-induced reasons for the RAI system's decisions. These tools can enable potentially negligent acts to be more effectively identified and analysed.

[62] The PCRC, in evaluating the possible attribution of liability on corporations under a new offence of "failure to prevent an offence", felt that the mens rea for that offence should be one of negligence, but that the offence should be accompanied by guidance in subsidiary legislation on what constitutes a "reasonable" standard of care. *PCRC Report*, above, n 43 at 217.

[63] See, for example, Arun Rai, "Explainable AI: from black box to glass box." (2020) 48 J. of the Acad. Mark. Sci. 141.

[64] The importance of explainability, and its implications for laws and regulations are discussed further in Law Reform Committee, Singapore Academy of Law, *Applying Ethical Principles for Artificial Intelligence and Autonomous Systems in Regulatory Reform* (July 2020) (Co-Chairs: Justice Kannan Ramesh and Charles Lim Aeng Cheng) <*https://www.sal.org.sg/Resources-Tools/Law-Reform/AI_Ethical_Principles*> (accessed 1 February 2021).

For comparable reasons, we also support the promotion of standards to ensure robust, transparent and replicable testing as possible further means to ameliorate some of these challenges.

2 Alternative criminal liability mechanisms in situations of non-intentional harm from RAI systems

4.39 In this section, we consider three other possible bases, beyond reliance on existing criminal negligence laws, on which criminal liability might be attributed to address non-intentional harms.

4.40 These are not intended to be exhaustive or mutually exclusive. Nor are we seeking to advocate that any such basis should be used in relation to a specific form of harm or specific RAI application – or indeed that criminal, rather than civil or regulatory sanctions are the most appropriate legal response in a given scenario. Rather we hope to provide a broad indication of differing ways in which criminal laws might be formulated and applied to non-intentional harms 'caused' by RAI systems, and the extent to which they may be capable of addressing the various challenges thus far identified in this report.

I Legal personality for RAI systems as a model?

4.41 One possible solution to some of these challenges that has been debated is the possibility of imposing liability directly on the RAI system itself. This would require the conferment of separate legal personality on an RAI system, similarly to the way legal personality is presently conferred on companies.[65]

4.42 For example, Jacob Turner, the author of *Robot Rules: Regulating Artificial Intelligence*, while acknowledging the challenges of how exactly such separate legal personality would be structured, has suggested that the possibility should not be discounted out of hand. Granting legal personality to an RAI entity, he asserts, would not necessarily mean treating it as a human, nor need it operate as a convenient mechanism for humans to disclaim responsibility for an RAI system's actions.[66]

[65] For example, the European Parliament, as part of a series of recommendations on robotics, has suggested that the European Commission "consider the implications of … creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause." *Civil Law Rules on Robotics: European Parliament Resolution of 16 February 2017 with Recommendations to the Commission on Civil Law Rules on Robotics* (2015/2103(INL)), European Parliament (16 February 2017) at 59 <*http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2017-0051+0+DOC+PDF+V0//EN*> (accessed 1 February 2021).

[66] Jacob Turner, *Robot Rules: Regulating Artificial Intelligence* (Springer, 2019) at 205.

4.43 Such a legal fiction may limit the extent to which it is necessary to get 'under the bonnet' of an RAI system, and identify which specific part or parts of that system caused its decision to act as it did, and which of the (potentially numerous) parties involved in the system's development and deployment should be held responsible there for.[67]

4.44 Further, as noted above, the notion of non-human entities being the responsible 'actor' in a crime is not without precedent. Corporate entities can carry out, and be held liable for, numerous criminal acts. Such liability has an indirect penalising effect[68] on the stakeholders/shareholders and the officers of the corporation. Thus, the argument runs that according legal personality and liability to an RAI system could have a similar indirect penalising effect on those responsible for or profiting from the RAI system, while minimising the need to prove that the harm was attributable to specific natural persons or corporations.

4.45 A key counter-argument – and one recently advanced by an expert group established by the European Commission – is that it is not in fact necessary to give devices or autonomous systems a legal personality to achieve the perceived benefits thereof. "Harm caused by even fully autonomous technologies," the expert group argued "is generally reducible to risks attributable to natural persons or existing categories of legal persons, and where this is not the case, new laws directed at individuals are a better response than creating a new category of legal person." [69]

4.46 Chesterman has posited a similar view. He notes the tendency of some arguments in favour of legal personality to take an anthropomorphised view of RAI systems (that is, focusing on humanoid robots), whereas in reality RAI systems "exist on a spectrum with blurred edges". As such, he argues, "there is as yet no meaningful category that could be identified for such recognition [of legal personality]". To the extent that there may, as with corporations, be "instrumental" reasons for conferring personality in specific cases, so as to avoid crimes going unpunished, Chesterman, like the European Commission expert group, believes that this could be achieved using "existing legal forms".[70]

[67] This could evidently be relevant in scenarios where either the human user-in-charge is not (or is only partially) liable, or where there is no direct human involvement at all when a particular incident occurs.

[68] In addition to any individual penalty imposed *directly* on such officers where the corporation committed the offence with their consent and connivance.

[69] See, Expert Group on Liability and New Technologies, New Technologies Formation, *Liability for Artificial Intelligence and Other Emerging Digital Technologies* (November 2019) <*https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=36608*> at 37-38 (accessed 1 February 2021). The Expert Group's recommendations followed the European Parliament's proposal that the implications of creating a specific legal status for robots be considered by the European Commission (see above, n 65).

[70] Simon Chesterman. "Artificial Intelligence and the Limits of Legal Personality" (2020) 69(4) International and Comparative Law Quarterly 819 at 843.

4.47 On balance, the arguments against separate personality for RAI systems appear more compelling, at least at the current stage of technological advancement.[71] We consider that criminal laws should continue to be formulated on the basis that such laws are intended to shape or impact *human* behaviour. It is not fully clear, for example, how imposing criminal liability (and a sanction) on an RAI system directly would deter the system *itself* from causing harm.[72] And to the extent that the objective would be to deter or penalise those responsible for the RAI system, rather than the system itself, we also take the view that such ends could equally be achieved through alternative mechanisms that do not require the creation of wholly new forms of legal personality (with all the disruption to the existing legal framework that that would necessarily entail). Indeed, it could be argued that the models discussed further below provide evidence of this. Finally, despite Turner's reassurances, we would also not lightly dismiss concerns that separate legal personality might enable those otherwise responsible to shift responsibility to the RAI personality.

II The offences considered by the PCRC for computer programs as a model?

4.48 An alternative approach was considered by the Penal Code Review Committee (‘**PCRC**’) in its 2018 Report.[73] In that report the PCRC, among other things, expressed concerns that computer programs could cause damage going beyond that which the conventional machinery envisaged by provisions such as section 287 PC was liable to cause.[74] This included non-physical harms such as deception, psychological distress, and economic loss.

[71] While the assessment *may* change once RAI systems move closer to being capable of ‘Artificial General Intelligence’ and are able to, for example, process the implications of their actions vis-à-vis themselves or others, there appears to be some consensus among AI experts that such technologies remain at least two decades away (if not more). Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, “When Will AI Exceed Human Performance?” (2018) 62 JAIR 729 at 731; Seth D Baum, Ben Goertzel, Ted G. Goertzel, “How Long Until Human-Level AI? Results from an Expert Assessment” (2011) 78(1) Technological Forecasting and Social Change 185; Philip Boucher, “How artificial intelligence works” (March 2020), European Parliamentary Research Service <*https://www.europarl.europa.eu/at-your-service/files/be-heard/religious-and-non-confessional-dialogue/events/en-20190319-how-artificial-intelligence-works.pdf*> (accessed 1 February 2021).

[72] By contrast, for corporations – which do have separate legal personality – the financial and reputational consequences of being found criminally liable impact human shareholders and officers, and deter them from future offences.

[73] *PCRC Report*, above, n 43 at 29-31.

[74] *Id.* at 29.

4.49 To address this, the PCRC set out – albeit merely as a "starting point for future discussions"[75] – two possible new criminal provisions that would, respectively:

- Target the creation of risk by developers or operators of computer programs through their rash or negligent creation, alteration or use of a computer program, even where no hurt or injury were caused; and
- Impose a duty on those with control over a computer program to take reasonable steps to cease harms that may result from computer programs after they manifest.

4.50 The first offence proposed by the PCRC is worded as follows:[76]

> *Whoever makes, alters or uses a computer program so rashly or negligently as to endanger human life, or to be likely to cause hurt or injury to any other person, or knowingly or negligently omits to take such order with any computer program under his care as is sufficient to guard against any probable danger to human life from such 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.*
>
> *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.*
>
> *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.*

4.51 The PCRC was also concerned that computer programs could cause serious harm in situations where the user did not know or intend that that would happen, and that, as we discussed above, such situations might not be adequately addressed by the Penal Code's existing provisions. It noted that one way to address this lacuna might be to impose a duty to take reasonable steps to cease such harms after they manifest. The PCRC therefore set out a second potential new offence that:[77]

> *Where a computer program —*
>
> *(a) produces any output, or*
>
> *(b) performs any function,*

[75] *Id.* at 30.

[76] *Ibid.* The PCRC noted that such an offence would be distinct from offences currently found under the CMA, which deals with cases where there is an intent to commit an offence, or a lack of authority to perform an act with a computer.

[77] *Id.* at 31—32.

> *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.*

4.52 We consider that the laws proposed by the PCRC may help to address two of the present challenges discussed, namely identifying the (legal) person(s) to whom liability should be attributed, and setting out the parameters of the duties to which such a person is subject.

4.53 Thus, under the PCRC's offences, liability could attach beyond human users to any creators, designers and/or corporations producing the RAI systems found to have negligently breached the standards required of them. An analogy from a non-RAI context might be the imposition of criminal liability on an architect or engineer whose negligence results in the collapse of a building and the loss of lives. Indeed, given the broad way in which the PCRC defines 'control' or 'care' over a computer program for these purposes (in essence, having 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),[78] the PCRC's second offence in particular may well extend to harm caused without any direct human involvement at all.

4.54 However – and understandably given their intended generality – the PCRC's offences do not stipulate the exact contours of the duties they impose. As such, they do not necessarily address the challenges of determining if a duty has been breached (i.e., what constitutes a rash or negligent act or failure to take reasonable steps in any given circumstance) – at least in the absence of soft-law guidance or the incremental development of relevant case precedent. As previously noted, such potential uncertainty as to the precise boundaries of the offences and the standards to which developers and others will be held could risk disincentivising them from deploying potentially beneficial RAI systems in Singapore, out of fear that they may be held criminally liable where harms are unintentionally caused.

4.55 Moreover, there remains the broader policy concern – highlighted at the outset of this report and acknowledged by the PCRC[79] – about the potential chilling effect of criminal liability for non-intentional harms on innovation in RAI systems in Singapore.

[78] *Id.* at 30, 32.

[79] In part for these reasons, the PCRC considered legislative change at this time to be "premature". It noted in particular that no other country had introduced specific rules on criminal liability for AI systems, and raised concerns about Singapore's ability to attract top industry players in the field of AI if it were to be the first to do so.

4.56 Even if an approach akin to that envisaged by the PCRC were to be adopted, therefore, it would appear to be prudent to limit any proposed offences to specific high-risk use cases and/or to set out more precisely and explicitly in legislation the relevant duties imposed – and, as applicable, standards to be met – in a given circumstance. This appears preferable, at this stage, to introducing criminal negligence offences that apply across all sectors and RAI system applications.

III Workplace safety legislation as a model?

4.57 There remains the possibility that the operation of an RAI system results in death, serious personal injury or widespread public harm, but no individual can be identified as having directly caused that harm (whether intentionally or negligently).

4.58 By definition, a negligence-based framework for criminal liability would be inapplicable in such cases, regardless of how precisely the relevant obligations or standards of conduct were set out. Yet there may still be strong public demands for someone to be held accountable. To narrow risks of such an "accountability gap", one approach could be to adopt a model more akin to that under the existing Workplace Safety and Health Act ('**WSHA**'), where a duty is imposed on specified entities to take, so far as is reasonably practicable, such measures as are necessary to avoid harm.[80, 81]

4.59 In the workplace context, those duties are imposed on occupiers and employers; in the context of RAI systems it might be whichever entity(ies) is best placed – on the basis of their 'proximity' to the RAI system and its operation, and their resources – to take action (i.e., to prevent, address and rectify dangers posed by RAI systems) and to change future outcomes.

4.60 It has been suggested (in the context of autonomous vehicles, although the same rationale would appear to apply equally to other forms of RAI system) that such proximity may arise in three ways:

[80] See the Workplace Safety and Health Act (Cap 354A, 2009 Rev Ed), s 12(1), which imposes criminal liability on employers to take necessary measures to ensure the safety and health of their employees at work in so far as reasonably practicable. As See Kee Oon J noted in *PP v GS Engineering and Construction Corp* [2017] 3 SLR 682 at [51], the introduction of more severe penalties for such offences in the Act was part of a concerted effort to deter poor safety management and effect a cultural change for employers and other stakeholders to take proactive measures to prevent accidents at the workplace. This demonstrates Parliament's intention to achieve such a deterrent effect and ensure that the true economic and social costs of such risks and accidents are borne by the responsible parties.

[81] For the avoidance of doubt, the burden of proof still rests with the prosecution to show a prima facie case of breach of duty. Once the prosecution satisfies that requirement, it is for the defendant to show that it took such reasonably practicable measures as were necessary to prevent the harm.

(a) the entity's proximity to the RAI system (insofar as it can exclusively access, diagnose and rectify an RAI system and its performance to ensure its proper functioning);

(b) its proximity to the user (insofar as it can shape how humans control or interact with the RAI system through the human-machine interface); and,

(c) its proximity to the task (insofar as it can unilaterally effect changes to the operation of the RAI system, for example via over-the-air-updates).[82]

4.61 In their ongoing inquiry into automated vehicles, the UK Commission applied similar 'proximity' principles in proposing that all self-driving systems should be backed by an Automated Driving System Entity (ADSE), being the entity that puts the vehicle forward for categorisation as 'safe self-driving' (meaning, in effect, as authorised to operate on public roads). Noting the "wide variety of organisations [that] may work together to develop self-driving vehicles", the UK Commission emphasised its belief in the importance of having a single, identified entity that would be "the first point of reference in the event of problems" and be subject to sanctions if things go wrong. That might be the vehicle manufacturer, a software developer or a partnership between the two. Crucially, however, the ADSE would need to show that it had been sufficiently involved in assessing safety and writing the safety case for the vehicle to vouch for the information in it, and have sufficient financial resources (e.g., to organise a recall or make mandated improvements where harms have arisen).[83]

4.62 Approaches akin to that in the WSHA shift the focus away from investigators having to determine the specific cause of the harm, or to prove negligence on the part of a (natural or legal person), and move instead towards a focus on whether the relevant entity breached its statutory duty to take all reasonably practicable measures to avoid the harm (including, for example, the adequacy of the protective processes and systems the entity had in place). A particular advantage of such approaches, from an enforcement perspective, is that the prosecution need not prove a direct or scientifically precise causation between the harm caused by the RAI and a particular breach of duty.

4.63 On the other hand, the sort of statutory duties imposed by the WSHA and similar legislation evidently place significant onus on defendants, and caution is therefore merited. Nonetheless, for policy reasons, such a burden may be considered to be justified in specific (and likely exceptional) circumstances, or for particular technologies, where (a) risks of serious harm are particularly acute or there was considered to be a particular

[82] Ella Pyman, "The Liability Blind Spot: Civil Liability's Blurred Vision of Conditionally Automated Vehicles" (2018) 92 ALJ 293, at 297—299.

[83] *UK AV Consultation*, above, n 32 at [8.64] – [8.67].

moral imperative on the entity in question to prevent the RAI system causing harm, and (b) there is a strong public desire for accountability.[84] Ultimately, whether such an approach is justified in any given context is a policy judgment for lawmakers, balancing, in particular, demands for accountability with the desire not to unduly stifle innovation and impede the societally-beneficial development and use of RAI systems.

4.64 That balance was also considered by the UK Commission. Having initially sought views on whether, where autonomous vehicles cause death or serious injury, new or stricter forms of criminal liability might be necessary or appropriate to avoid an accountability gap, the UK Commission appears to have adopted a relatively narrow approach to the use of criminal liability. Specifically, the recommendation on which it is now consulting is that:

(a) The primary means to promote safety would be a system of (non-criminal) regulatory sanctions on ADSEs, with the focus on identifying and addressing the problem to prevent recurrence, rather than on blame.[85]

(b) Rather than leaving it to system developers to determine 'how safe is safe enough' on the understanding that when things go wrong they would bear the responsibility, it was preferable to establish a "relatively tight" safety assurance system, coupled with third party testing. If ADSEs comply with that assurance and testing scheme and act honestly, they would not be blamed for adverse outcomes.[86]

(c) To the extent that new criminal offences were necessary, these should complement the safety assurance regime and be limited to situations of "serious wrongdoing" where an ADSE failed to provide information or misled the safety regulator (i.e. there is an element of 'dishonesty'). In those circumstances, it would be the ADSE as a corporate body, that was held liable, subject to a due diligence defence.[87] It was

[84] In areas such as workplace safety, where distinct statutory duties are already imposed and the appropriate policy balance has thus already been considered and determined, it may evidently not be necessary to create a wholly new legal framework specifically for RAI systems. Rather, it would seem prudent to instead review (and as necessary amend) those existing laws to ensure that occupiers and employers may equally be held responsible for harm resulting from the autonomous operation of RAI systems in their workplaces (subject to any available defences in the WSHA or equivalent legislation).

[85] *UK AV Consultation*, above, n 32 at [14.1].

[86] *Id.* at [14.13] – [14.14].

[87] *Id.* at [14.5] and [14.18] – [14.20]. Senior managers of the entity might also be liable, if the offence was committed with their consent or connivance, or was attributable to their neglect.

emphasised that such offences were "different from criminalising negligence."[88]

IV Targeting and proportionality

4.65 As we have noted, regardless of the model adopted, one means to mitigate concerns about the potential chilling effect on innovation or the imposition of an unduly onerous burden on an entity is to target offences to particular high-risk sectors or use cases where, for example, the potential for serious harm is greatest, or the need for deterrence strongest.[89] Additionally, to further mitigate those concerns, the sanctions imposed on the entity could be calibrated to ensure that they are proportionate to the nature of the offence and the entity's degree of blameworthiness. This is already the case with criminal negligence offences (where the statutory penalties are generally lower than for comparable offences committed intentionally), and the Singapore courts have, for their part, used both culpability and harm caused as analytical bases for sentencing benchmarks that they have issued.[90]

4.66 While it is important that the law does not serve inadvertently to deter benign development and deployment of RAI systems, it is equally the case that that law should promote high standards of safety, particularly where risks are greater. A well-targeted criminal liability regime offers an effective means to strike that balance and one which reflects the seriousness of the potential harm that RAI systems may cause if necessary standards are not maintained.

4.67 Of course, there may well be reasons why criminal liability may not be considered necessary or appropriate, either in a particular sector, in relation to a particular act or omission, or more generally. In such circumstances, models akin to those discussed above might still be suitable for adoption, but (as the UK Commission's consultation proposes)

[88] *Id.* at [14.18].

[89] In this regard, we note that a recent European Commission White Paper proposed the creation of a regulatory framework for AI under which greater regulatory requirements (regarding, e.g., the quality of training data, data and record-keeping, disclosure duties and human oversight, etc.) would apply to those AI applications considered to be 'high-risk'. 'High-risk' applications would be those meeting two cumulative criteria: (i) the AI application is employed in a sector where, "given the characteristics of the activities typically undertaken, significant risks can be expected to occur"; and (ii) the AI application is, in addition, used "in such a manner that significant risks are likely to arise". European Commission, *White Paper on AI: a European approach to excellence and trust* COM(2020) 65 (February 2020) <*https://ec.europa.eu/info/files/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en*> (accessed 1 February 2021).

[90] See Shawn Ho His Ming, "Analytical Framework for Sentencing: Harm, Culpability and Aggravating/Mitigating Factors", Singapore Law Gazette (January 2019) <*https://lawgazette.com.sg/feature/analytical-framework-for-sentencing/*> (accessed 1 February 2021).

reframed so as to impose only regulatory controls – the principal objective being to ensure that the relevant entities take all necessary measures to ensure that their RAI systems continue to operate normally and safely. The enforcement of such regulatory controls need not be by way of criminal sanctions. Non-criminal sanctions such as civil financial penalties and revocation or suspension of licences might be considered sufficient to promote the development of safe RAI systems, and their safe deployment.

4.68 Finally, we note for completeness that, while we have evaluated the various models and approaches in this report for their effectiveness in addressing some of the challenges with attributing liability where the operation of RAI systems results in harm, we have not sought to explore in detail how they could be implemented (whether through legislation or other means). Rather, we consider that, at this early stage, it is preferable to articulate a broad framework through which the concepts and issues at play can be objectively analysed, and for this to provide a starting basis for future detailed analysis and debate on whether, when and how criminal law should apply in relation to the actions of autonomous RAI systems.

CHAPTER 5

CONCLUSION

5.1 Notwithstanding the immense benefits that they can bring, the risks and challenges posed by RAI systems – and their use in an ever-growing number of domains – heighten the need to ensure that the law effectively protects society from these risks and promotes the development of safe, societally-beneficial RAI technologies. That is particularly the case in relation to applications of RAI that are likely to be used in a way, or context, in which risks of harm are significant.

5.2 This report has explored the challenges that arise in seeking to apply existing criminal laws and principles in relation to the actions of increasingly autonomous RAI systems, and considered ways in which those laws or principles might need to be adapted, or new laws or approaches adopted. All approaches have their strengths and weaknesses, and – given the wide variety of RAI systems and the myriad applications and settings in which they may be used – a 'one size fits all' approach is unlikely to be practicable or appropriate. And as we have noted throughout the report, regardless of the specific changes adopted, any new criminal laws should be proportionate, and the imposition of criminal liability and sanctions carefully calibrated to avoid penalising benign conduct or deterring beneficial activity.

5.3 What remains clear, however, is that the use of RAI technologies will continue to give rise to new forms of harm and thus to continue to challenge existing laws and regulations, requiring legislators to respond with agility to new and emergent risks. We hope that this report will be a catalyst to proactive analysis and well-informed debate, even as the technology rapidly evolves.

GLOSSARY[91]

Adversarial AI — the malicious use of models or signals that, while typically resembling normal inputs, are designed to cause an **AI system** to misinterpret that input and 'fool' it into behaving in erroneous ways.

AI System — a machine-based system able, for a given set of human-defined objectives, to make predictions, recommendations, or decisions that influence real or virtual environments. Such systems are able to operate with some level of **autonomy**, and can be incorporated into hardware devices or entirely software-based.

Algorithm — a set of rules or instructions (i.e. mathematical formulas and/or programming commands) given to a computer for it to complete a given task.

Artificial Intelligence (AI) — a set of technologies that seek to simulate human traits such as knowledge, reasoning, problem solving, perception, learning and planning, and, depending on the AI model, produce an output or decision (such as a prediction, recommendation, and/or classification).[92]

Auditability — the readiness of an **AI system** to undergo an assessment, by internal or external auditors, of its **algorithms**, **data** and design processes.

Autonomy/autonomous — the ability of an **AI system** to function (i.e. to take decisions and act) independently without human intervention.

Bias — the distortion or skewing of an **AI system**'s outputs, either due to the design of the algorithm or due to the input **datasets** utilised by the AI system being unrepresentative or discriminatory. Two common forms of bias in data include:

- selection bias (when the **data** on which an **AI system** bases its outputs are not representative of the actual **data** or environment in which the **AI system** operates); and
- measurement bias (when the process or means by which **data** is collected results in that gathered **data** being skewed or distorted).

[91] The definitions in this glossary have been adapted from various sources for the specific purposes of the present series of reports. They are intended as an aid to the reader and should not be treated as exhaustive or authoritative.

[92] We note that there is no widely-accepted or authoritative definition of artificial intelligence. The definition used here is a non-exhaustive, adapted definition used in the *Model AI Governance Framework (Second Edition)*, above, n 26.

Big Data — datasets characterised by their:

- size ("Volume");
- complexity ("Variety") (i.e. typically including structured, semi-structured and unstructured data derived from diverse sources); and/or,
- rate of growth ("Velocity"),

from which detailed insights can be derived using advanced analytical methods and technologies (e.g., **neural networks** and **deep learning**).

Black box (1) —an **AI system** whose decision-making operations are not **explainable** – that is, the means by which it reached a particular decision or action are neither disclosed nor able to be ascertained by human **users** or other interested parties (for example regulators, testers or auditors).

Black box (2) — see **Event Data Recorder**.

Bot — a software program (typically operating on the internet) designed to run automated tasks.

Chatbot — an **AI system**, commonly used in customer-facing commercial settings, designed to engage in dialogue with a human **user** via voice or written methods, and thus to simulate a human-to-human conversation. As the Chatbot engages in more conversations, it learns to better respond to future questions and more closely imitate real conversations. Examples include the "Ask Jamie" chatbot on the Singapore Ministry of Health's website, or the 'Live Chat' help functions on e-commerce platforms such as Lazada or Shopee.

Cyberattacks — a malicious attack launched from one or more computers against other computers, networks or devices.

Data — information defined as and stored in code to be processed or analysed. Individual records of data (for example a person's name or the temperature recorded by a smart home device at a particular date and time) can be combined together to form **datasets**. A distinction is commonly drawn between personal data (those which individually or in combination with other data, identify an individual) and non-personal data (those that do not).

Data portability — the legal obligation to comply with a data subject's request for their **data** to be moved from one organisation to another in a commonly used machine-readable format.

Dataset — a collection of data (often stored in the form of one or more databases).

Deep learning — a specific form of **machine learning** that utilises **neural networks** to model and draw insights from complex structures and relationships between **data** and **datasets**. The term derives from the ‘layers’ of the **neural network** down through which the **data** passes.

Deployer — the person or legal entity responsible for putting an **AI system** on the market or otherwise making it available to users. The deployer may also have on ongoing role in operating or managing the **AI system** after deployment.

Derived data — any **data** element that is created and/or derived by an organisation through the processing of other **data** in the possession and/or control of the organisation.

Designer / Developer — a person or legal entity who takes decisions that determine and control the course or manner of the development of **AI systems** and related technologies. ‘Development’ for these purposes means (a) designing and constructing **algorithms**, (b) writing and designing software, and/or (c) collecting, storing and managing **data** for use in creating or training **AI systems**.

Event Data Recorder — a machine that continuously records the inputs received by an **AI system** (e.g. what its sensors ‘see’), its relevant internal status data, and its outputs. Sometimes colloquially known as a ‘black box recorder’. The intention of such event data recorders, equivalent to those installed in aircraft, is to allow post-hoc analysis of the **AI system**’s operation (e.g., in the lead up to an accident or system failure).

Explainability — the ability for a human, by analysing an **AI system**, to understand how and why the system reached a particular decision or output.

Explainable AI — broadly, either (a) AI systems which are designed to be inherently **explainable**, such that a human can understand how and why the system reached a particular decision or output; or (b) tools designed to help extract explanation from pre-existing **black box** and other complex **AI systems**.

Human-Machine Interface — a screen, dashboard or other interface which enables a human **user** to engage with an **AI system** or other machine.

Internet of Things, the (IoT) — a system comprised of interconnected devices (commonly known as smart devices) that transfer **data** and communicate with one another via the internet.

Machine Learning — a technique whereby a set of **algorithms** utilise input **data** to make decisions or predictions, and thus to ‘learn’ how to complete a task without having been specifically programmed to do so.

(Artificial) Neural Networks — a series of ‘layered’ **algorithms** used to analyse, classify, learn from and interpret input **data**. The values from one layer are fed into the next layer to derive increasingly refined insights.

Artificial Neural Networks are so named because they broadly mimic the biological neural networks in the human brain.

Operator — see **User**.

Robotics — technologies that enable machines to perform tasks traditionally performed by humans, including by way of **AI** or other related technologies. This series of reports focuses on robots that act fully or partially autonomously, without human intervention.

Robustness — the ability of an **AI system** to deal with errors that arise during execution or erroneous input, and to continue to function as intended or without insensible, unexpected or potentially harmful results.

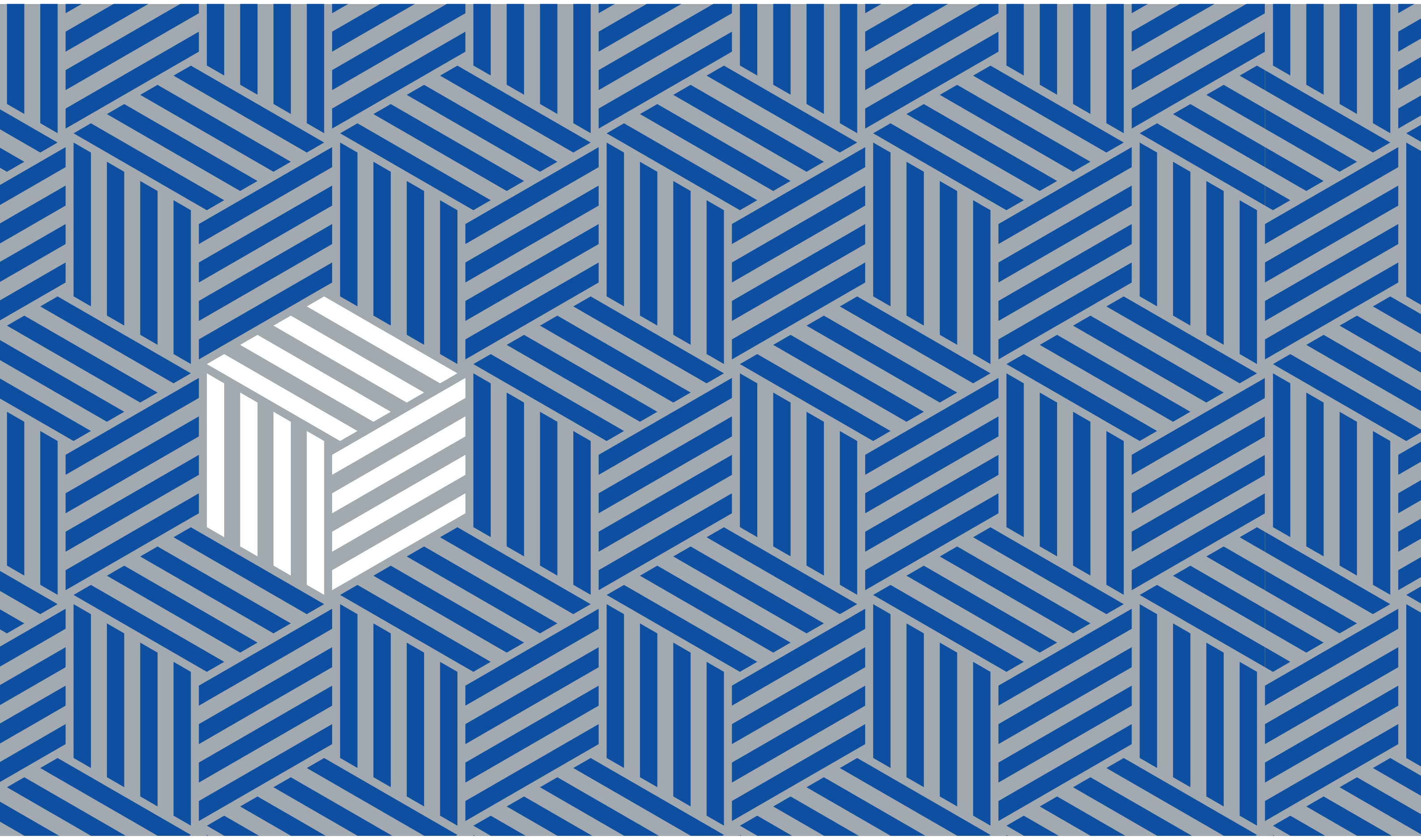
SAE Levels — a classification system developed by the Society of Automotive Engineers International, which classifies **autonomous** vehicle technologies according to six levels of increasing automation (and declining human involvement).

Traceability — the documentation, in an easily understandable way, of (a) an **AI system**'s decisions, and (b) the **datasets** and processes that yield those decisions (including those of data gathering, data labelling and the **algorithms** used). This provides a means to verify the history, and contexts in which decisions are made.

Transparency — various mechanisms or requirements intended to provide additional information to users, regulators and other stakeholders regarding the algorithmic decision-making processes undertaken by **AI systems**, and the input **data** relied on by such systems. Such transparency may be achieved through, for example, disclosure of source code, **explainability** and/or **traceability**. Transparency also implies that **AI systems** should (in practice, and by design) carry out their functions in the way communicated to others (including **users**).

User — any natural or legal person who uses an **AI system** for purposes other than development or deployment.

Verifiability — the process of ensuring that the outputs of an **AI system** correspond with its intended function or purpose (for example by testing the system using a range of different inputs, or ensuring that a particular input consistently and repeatedly leads to a desired output).



ISBN 978-981-18-0164-8 (softcover)
978-981-18-0165-5 (e-book)