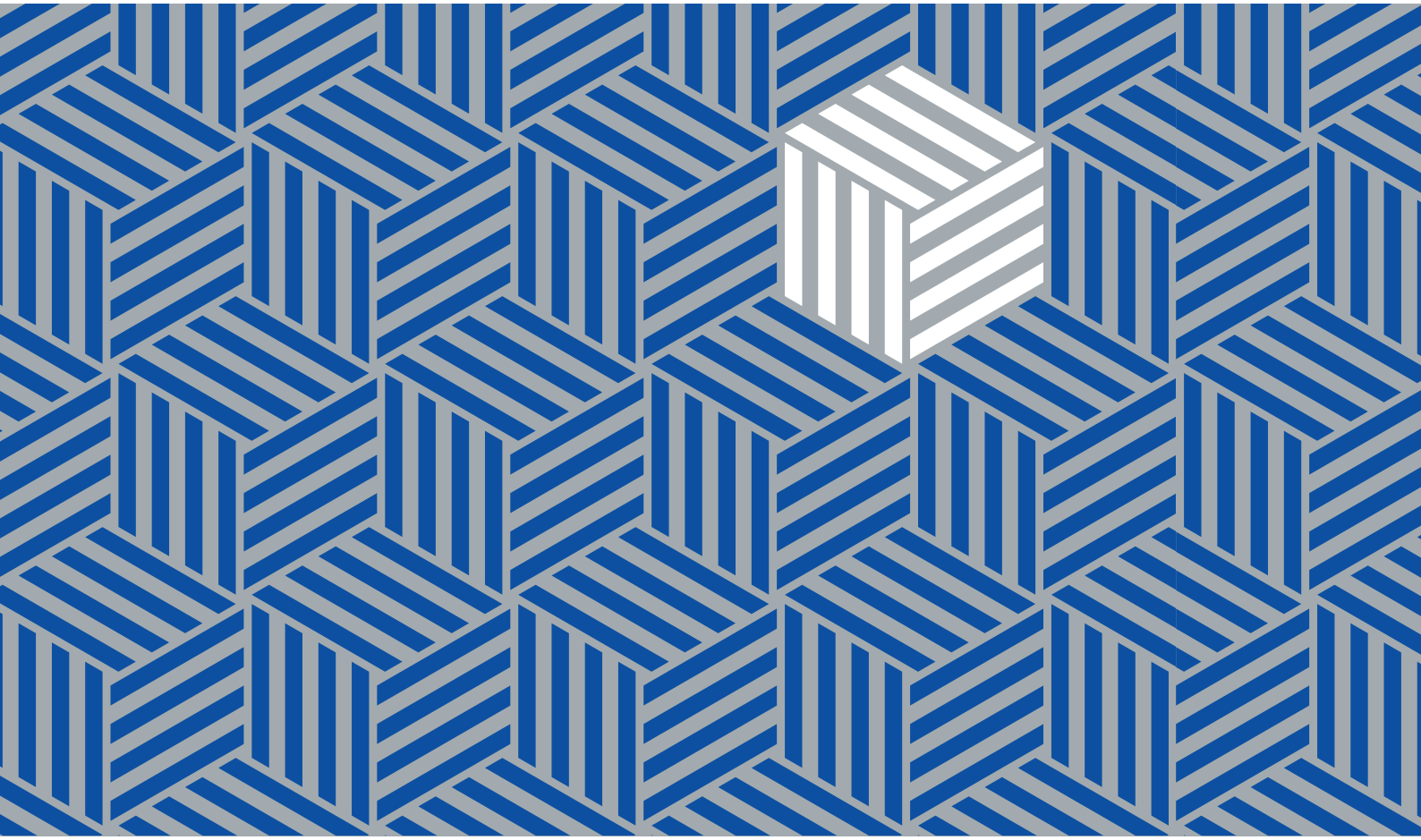
Singapore Academy of Law
Law Reform Committee

Applying Ethical Principles for Artificial Intelligence in Regulatory Reform

July 2020



SAL SINGAPORE ACADEMY OF LAW

Singapore Academy of Law
Law Reform Committee

Applying Ethical Principles for Artificial Intelligence in Regulatory Reform

July 2020

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

SAL SINGAPORE ACADEMY OF LAW

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

National Library Board, Singapore Cataloguing in Publication Data

Name(s): Singapore Academy of Law. Law Reform Committee. | Constantine, Simon, editor.
Title: Applying ethical principles for artificial intelligence and autonomous systems in regulatory reform / Singapore Academy of Law, Law Reform Committee; edited by Simon Constantine.
Other title(s): Impact of robotics and artificial intelligence on the law
Description: Singapore: Law Reform Committee, Singapore Academy of Law, [2020]
Identifier(s): OCN 1158290995 | ISBN 978-981-14-6600-7 (paperback) | ISBN 978-981-14-6601-4 (ebook)
Subject(s): LCSH: Artificial intelligence–Law and legislation–Singapore. | Robotics–Law and legislation–Singapore. | Artificial intelligence–Moral and ethical aspects–Law and legislation–Singapore. | Robotics–Moral and ethical aspects–Law and legislation–Singapore.
Classification: DDC 343.59570999–dc23

ISBN 978-981-14-6600-7 (softcover)
978-981-14-6601-4 (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 ranked first in the 2019 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 its 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

- Rethinking Database Rights and Data Ownership in an AI World (published July 2020)
- Report on the Application of Criminal Law to the Operation of Artificial Intelligence Systems and Technologies (*forthcoming*)
- Report on the Attribution of Civil Liability for Accidents Involving Automated Cars (*forthcoming*)

TABLE OF CONTENTS

APPLYING ETHICAL PRINCIPLES FOR ARTIFICIAL INTELLIGENCE IN REGULATORY REFORM

EXECUTIVE SUMMARY

1 It is clear that Artificial Intelligence (AI) systems and other autonomous technologies, by reducing the need for human intervention in a range of processes, have the potential to offer significant human and societal benefits, but also to create risks which should be guarded against.

2 It is ultimately for policymakers to determine the range of interventions necessary across different sectors to achieve the optimal balance between encouraging innovation and the development and deployment of societally-beneficial AI systems on one hand, and protecting and promoting human wellbeing on the other. That assessment will require consideration and balancing of numerous factors, and where the balance lies may vary between technologies or sectors.

3 However, there is an increasing – and in our view welcome – acceptance that the deployment of AI should be underpinned by an ethical framework that helps to ensure that those technologies improve human wellbeing. Various governments and non-governmental organisations have already put forward principles, recommendations and the like on AI ethics and governance. Many of these take the form of voluntary principles directed at those developing or deploying AI systems.

4 Accordingly, this report considers the core ethical principles for which there is emerging consensus, and assesses their implications, and the challenges they may raise, for policymakers in formulating any hard or soft law interventions considered necessary.

5 We hope this report will also assist other relevant stakeholders – including AI system designers, deployers and users – in understanding the principles and practices that they can adopt as voluntary best practices to help promote the ethical and societally-beneficial development and adoption of AI systems.

6 The following ethical principles are analysed: a) respecting fundamental interests; b) considering effects; c) wellbeing and safety; d) managing risks to human wellbeing; e) respect for values and culture; f) transparency; g) accountability; and h) the ethical use of data.

7 The implications and challenges for lawmakers of each principle vary, and it is clear that there is no 'one size fits all' regulatory solution, not least given the diversity of AI systems (from mobile phone apps to autonomous vehicles to sophisticated 'affective' robots designed to mimic human emotions) and the contexts in which they may be deployed.

Different regulatory tools and different levels of intervention may be appropriate for addressing different ethical principles and for the different contexts in which such principles arise.

8 This report does not seek to advocate any specific means or level of intervention. Rather, it is hoped that, by identifying issues policymakers may encounter, the report will contribute to the consideration of how ethical principles can be incorporated into the development of fair, just and consistent laws, regulations and 'soft law' measures that foster technological development that prioritises human wellbeing and promotes human dignity and autonomy.

9 It is hoped that an international consensus on principles for a common ethical framework will emerge, and that this will, in turn, compel and incentivise the responsible development and deployment of AI systems across the world. By developing our own framework and human-centred approach to AI, Singapore would have contributed in a small way to developing this international consensus.

CHAPTER 1

INTRODUCTION

1.1 In the past few years, research and development in, and design and implementation of, artificial intelligence[1] ('**AI**') and autonomous systems and technologies ("**AI systems**")[2] have advanced by leaps and bounds.

1.2 Specific activities involving AI may be driven, in the first instance, by private interests. However, when viewed from a wider perspective, society as a whole benefits from advances in AI systems that reduce the need for human intervention in a growing range of processes. Such benefit could be in the form of greater efficiency in some areas, or less human error in others. Yet increased automation will also disrupt many aspects of society, require adjustments to be made and pose potential risks.

1.3 At a policy level, these developments therefore require a multi-dimensional assessment aimed ultimately at achieving the best balance between two goals (which may sometimes pull in different directions): encouraging innovation and fostering the research, development and deployment of societally-beneficial AI systems on one hand, and protecting and promoting human wellbeing on the other.[3]

1.4 Such balancing is ultimately a decision for policymakers to make, following an assessment of the costs, benefits, risks and other factors at issue in a given case, and taking into account, in particular, the rapidly evolving landscape and the need for any intervention to be sufficiently adaptable and responsive thereto. As many of the relevant technologies may, in effect, operate across borders,[4] international norms or standards may also be of particular relevance.

---

1 The definition of AI is not without contention. For this paper, we adopt the non-exhaustive definition used in the Singapore Infocomm Media Development Authority ("**IMDA**") and Personal Data Protection Commission Singapore ("**PDPC**")'s Model Artificial Intelligence Governance Framework, Second Edition (2020) ("**Model AI Governance Framework**") at 18. <*https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf*> (Accessed 10 June 2020).

2 In this chapter, AI systems include both AI software able to change its own outputs or programs based on data, and also systems that incorporate AI software as a component.

3 See further, Model AI Governance Framework at 8, and United States Executive Office of the President, Office of Management and Budget, Draft Memorandum to the Heads of Executive Departments and Agencies, "Guidance for Regulation of Artificial Intelligence Applications" (13 January 2020) ("**OMB Draft Guidance for Regulation**") at 2 <*https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf*> (accessed 10 June 2020).

4 For example, an AI system based in one country may use input data gathered in another and be overseen by a human operator situated in a third.

1.5 As in a non-AI context, where regulatory intervention is considered appropriate, there are likely to be a variety of approaches available:

- These may include legislation, subsidiary legislation, general or sector-specific guidance or frameworks, regulatory sandboxes, supervisory or monitoring authorities, or the development of consensus codes of practice, standards or certification mechanisms.
- Such interventions need not necessarily involve the imposition of additional requirements or controls on the creation or use of AI systems. They might equally be targeted at removing impediments to the development of such systems or at refining existing regulatory mechanisms.

1.6 However, there has been an increasing acceptance internationally that – given the profound changes widespread deployment of AI and autonomous technologies will precipitate – such deployment needs to be underpinned by an ethical framework that helps ensure those technologies improve human wellbeing.

1.7 To that end, various governments and non-governmental organisations have put forward principles, frameworks and recommendations on AI ethics and governance. Primarily, these have been through the formulation of voluntary principles for those developing or deploying AI systems. Locally, for example:

- The Singapore Government has recently published the second edition of its Model AI Governance Framework. This framework provides practical guidance to private sector entities on key ethical and governance issues when deploying AI solutions. It is founded on two core principles: (i) AI decision-making processes should be explainable, transparent and fair, and (ii) AI solutions should be human-centric.[5]
- Similarly, in the financial services context, the Monetary Authority of Singapore ("MAS") has used its previous development of various principles to promote fairness, ethics, accountability and transparency in AI use[6] as the foundation for the Veritas initiative, a framework for promoting responsible adoption of AI and data analytics. The first phase of that initiative involves the development of metrics in credit risk scoring and customer marketing, leading to the release of

5 Model AI Governance Framework, above, n 1 at [2.7].

6 Monetary Authority of Singapore, "Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector" (7 February 2019) ("**FEAT Principles**") <*https://www.mas.gov.sg/publications/monographs-or-information-paper/2018/FEAT*> (accessed 10 June 2020).

open source code which may be integrated in the IT systems of financial institutions.[7]

1.8 In this report, we examine the broad ethical principles[8] for which there is emerging consensus,[9] and consider their implications for policymakers in promulgating hard or soft law interventions in circumstances where regulatory intervention is considered necessary. For each, examples are given of areas where challenges may arise, and/or of questions that policymakers will likely need to consider. The report's analysis is intended to be technology-neutral and capable of application and adaptation by stakeholders in a manner appropriate to the nature of the specific technology, application, industry or risks at issue.

1.9 The primary objective of this report is to advance a public discussion about how those ethical principles – which may not be readily-measurable in quantitative terms – can be incorporated into the development of **fair, just, appropriate and consistent laws, regulations and 'soft law' measures that foster technological development that prioritises human wellbeing and promotes human dignity and autonomy**.[10] It does not seek to suggest or recommend specific legal or regulatory reforms, nor to suggest how complexities should be resolved.

---

7 MAS, "'Fairness Metrics' to Aid Responsible AI Adoption in Financial Services" (28 May 2020): <*https://www.mas.gov.sg/news/media-releases/2020/fairness-metrics-to-aid-responsible-ai-adoption-in-financial-services*>; MAS, "MAS Partners Financial Industry to Create Framework for Responsible Use of AI" (13 November 2019): <*https://www.mas.gov.sg/news/media-releases/2019/mas-partners-financial-industry-to-create-framework-for-responsible-use-of-ai*> (accessed 10 June 2020).

8 See for example: the Model AI Governance Framework, above n 1 at Annex A; the "Code of Ethical Conduct for Robotics Engineers" in European Parliament, "Report with Recommendations to the Commission on Civil Law Rules on Robotics" (2015/2103(INL)), A8-0005/2017 (27 January 2017) ('**Civil Law Rules on Robotics**') at 22 <*https://www.europarl.europa.eu/doceo/document/A-8-2017-0005_EN.pdf*> (accessed 10 June 2020); European Parliament, "Draft Framework of ethical aspects of artificial intelligence, robotics and related technologies" (2020/2012(INL)) (21 April 2020) ("**Draft Framework of Ethical Aspects**") <*https://www.europarl.europa.eu/doceo/document/JURI-PR-650508_EN.pdf*> (accessed 10 June 2020); Australian Department of Industry, Science, Energy and Resources, "AI Ethics Principles" (2019) <*https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles*> (accessed 10 June 2020); and industry-specification guidelines such as the Monetary Authority of Singapore's 'FEAT Principles', above, n 6.

9 As the Model AI Governance Framework notes "There are already a number of attempts globally in establishing a universal set of principles. While a consistent core set of ethical principles is emerging, there is also a penumbra of variation across cultures, jurisdictions and industry sectors.", above n 1 at [1.2].

10 Compare the Draft AI R&D Guidelines for International Discussions (Tokyo: The Conference toward AI Network Society, 28 July 2017) at [2.1] ('**Draft AI R&D Guidelines**'), available on the website of the Ministry of Internal Affairs and Communications of Japan at <*http://www.soumu.go.jp/main_content/000507517.pdf*> (accessed 10 June 2020); Draft Framework of Ethical Aspects, above n 8, at 5.

1.10 We hope this report will also assist other relevant stakeholders – including AI system designers, deployers and users – in understanding the principles and practices that they can adopt as voluntary best practices to help promote the ethical and societally-beneficial development and adoption of AI systems.

1.11 We consider that public discussion of these issues would benefit from the input of a wide range of relevant stakeholders, including governments, private corporations, civil society, individual citizens, and educational institutions. In particular, we believe that a high level of public consultation and participation in the establishment of norms in this space would help engender public trust and thus wider adoption of AI.[11]

1.12 Some of the issues raised in this report will be considered further in the subsequent reports in this Series. Those reports will analyse specific areas of law where the Committee considers that reforms to adapt to the development and widespread adoption of AI systems may be beneficial or necessary.

11 OMB Draft Guidance for Regulation, above, n 3 at 3.

CHAPTER 2

THE ETHICAL PRINCIPLES

A. LAW AND FUNDAMENTAL INTERESTS

AI systems should be designed and deployed to comply with law and not violate established fundamental interests of persons protected by law.[12]

2.1 AI systems may be vulnerable to misuse. Hackers may exploit weaknesses in AI's security systems to carry out cyberattacks or cause other harms.[13] As more AI systems are able to pass the Turing Test (that is, a human evaluator is unable to distinguish those systems' responses from a human's), some may use AI systems to pose as humans for unethical or criminal activities.

2.2 It is therefore imperative that there are effective legal or other safeguards that proscribe such activities and prevent or mitigate such abuse.[14] Effectiveness in this context means that those safeguards both:

(a) ensure that those responsible for harms are held responsible; and also

(b) 'make sense' when applied or extended to AI systems, and do not serve to preclude benign behaviour or to impose a level of compliance costs that deter beneficial uses of AI.

2.3 However, questions of criminal or civil legal liability may not be straightforward with AI systems. While it may be fairly easy to identify a wrongful act or effect resulting from the use of an AI system, it will often be less straightforward to identify the blameworthy actors.

- Laws imposing criminal or civil liability typically have regard to the mental state of the relevant actors (such as knowledge or intention), yet AI systems have no such mental state (at least not in the sense that term has been historically conceived).

12 These include, for example, the rights of persons to life, freedom from bodily harm and wrongful restraint, and protection of property, as exemplified by crimes against property listed in the Penal Code (Cap 224, 2008 Rev Ed), and protection of one's reputation in the form of the common law tort of defamation.

13 UK House of Lords Select Committee on Artificial Intelligence, "AI in the UK: ready, willing and able?", Report of Session 2017-2019 (16 April 2018) at [319]-[329] ("**UKHL 2018 AI Report**"): <*https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf*> (accessed 10 June 2020).

14 Draft AI R&D Guidelines, above, n 10 at [4.5] and [5.5]; Draft Framework of Ethical Aspects, above n 8, at 5.

- Furthermore, a 'decision' by an AI system to act is the result of a long causation chain involving different actors at different stages of the system's creation and deployment. This contrasts with a human's individual cognitive choice to carry out a particular action, and further complicates the task of identifying blameworthy actors.

2.4 Where a person *intentionally* uses an AI system to commit wrongful acts, saying that that person should be held liable would appear uncontroversial and unproblematic. However, what if a person's use of an AI system results in wrongful acts that are 'unintended' in that they are the consequence of rash or negligent conduct? In such a context, policymakers drafting new laws or amending existing ones may need to consider a range of potentially complex questions. These may include, for example:

(a) When assessing liability or associated standards of care, should there be an objective assessment of whether it was reasonably foreseeable that the wrongful act in question may result from the use of the AI system?[15]

If so, what is the level of specificity of the wrongful act which ought to have been reasonably foreseen?

For example, what if an AI system was designed and deployed to randomly purchase items on the Dark Web when the designer or deployer ought to have known that it could foreseeably purchase prohibited items such as drugs or weapons.[16] Or what if an AI bot trawls information on the web to write and publish a news report about a public figure – is it reasonably foreseeable that the bot would find and republish content that is defamatory?

(b) To what extent is the AI system's level and scope of automation relevant to the factual inquiry? If it is, would this raise significant practical difficulty in satisfying the relevant evidential burden?

For example, what if an AI system, in pursuing its programmed objectives, autonomously generated a bot which wrongfully accessed data in third party computer systems – should the designers or deployers be held responsible for breaching

15 Sections 26E and 26F of the Penal Code (Cap. 224) define "rashly" and "negligently" with reference to unreasonableness in acting while having knowledge of a real risk and reasonableness respectively.

16 Katie Grant, "Random Darknet Shopper: Exhibition Featuring Automated Dark Web Purchases Opens in London", Independent, 12 December 2015 <*https://www.independent.co.uk/life-style/gadgets-and-tech/news/random-darknet-shopper-exhibition-featuring-automated-dark-web-purchases-opens-in-london-a6770316.html*> (accessed 10 June 2020).

cybersecurity laws?[17] What if the aforesaid bot tasked to write a news report about a public figure autonomously generated a way to access personal information held in that person's private computer?

(c) How should the law address or assess the situation of a person who deploys a highly automated AI system, and that system then carries out a wrongful act which does not accord with the intentions of that person?

(d) Should a designer or deployer be held responsible for wrongful acts or effects caused by an AI system being subject to a malicious security breach, if such vulnerability was, or ought to have been, known?[18]

(e) Even if the harms caused by the AI system could not be foreseen or pre-emptively avoided, should there be duties imposed on a person to take reasonable steps to cease such harms after they manifest?[19]

2.5 Suggestions for beginning to address such issues have included, for example, imposing criminal liability on the person "who owns, operates or benefits from such a device or tool of technology",[20] or on the developer or deployer of an AI system who either "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."[21]

2.6 As stated at the outset, it is not within the scope of this paper to engage in a detailed analysis of the above issues, save as to raise them for careful consideration and to highlight their interrelation with core ethical principles. However, questions regarding the attribution of a) criminal liability for harms caused or facilitated by AI systems and autonomous

17 C.f. section 3(1) of the Computer Misuse Act (Cap. 50A): "Subject to subsection (2), any person who *knowingly* causes a *computer to perform any function* for the purpose of securing access without authority to any program or data held in any computer shall be guilty of an offence" [emphasis added].

18 Alexander Polyakov, "How AI-Driven Systems Can Be Hacked", Forbes 20 February 2018: <*https://www.forbes.com/sites/forbestechcouncil/2018/02/20/how-ai-driven-systems-can-be-hacked/#25e50d6079df*> (accessed 10 June 2020).

19 See, for example, *Report of the Penal Code Review Committee* (April 2018) at 31-32 <*https://www.mha.gov.sg/docs/default-source/default-document-library/penal-code-review-committee-report3d9709ea6f13421b92d3ef8af69a4ad0.pdf*> (accessed 10 June 2020).

20 The Senior Minister of State for Communications and Information and Education (Dr Janil Puthucheary), "Introduction of Regulations with Advent of Artificial Intelligence and Autonomous Machines" (6 February 2018), Par. No. 13, Session No. 1, Vol. 94, Sitting No. 59.

21 Report of the Penal Code Review Committee, above, n 19 at 30.

technologies;[22] and b) civil liability following accidents involving autonomous cars,[23] and possible means to address these will be considered by this Subcommittee in two forthcoming reports in this Series.

2.7 However, beyond mere avoidance of injury or harm, consideration might also be given to mechanisms that require those who develop and/or deploy AI systems to actively consider the effects of such development or deployment on the *fundamental interests* of persons and on nature and the environment.[24]

2.8 Precedents for such a principle can be found in existing codes of conduct, such as those adopted by the British Computer Society[25] ("*Members shall in their professional practice have regard to basic human rights and shall avoid any actions that adversely affect such rights.*") and the Association for Computing Machinery (ACM) ("*computing professional[s] should […] [c]ontribute to society and to human well-being, acknowledging that all people are stakeholders in computing. […] [including] promoting fundamental human rights*"). [26]

2.9 Similarly, the European Parliament Committee on Legal Affairs has previously proposed, as part of a Code of Ethical Conduct for Robotics Engineers, that:[27]

> *Robotics research activities should respect fundamental rights and be conducted in the interests of the well-being and self-determination of the individual and society at large in their design, implementation, dissemination and use. Human dignity and autonomy – both physical and psychological – is always to be respected […]*
>
> *Robot designers should consider and respect people's […] rights. A robotics engineer must preserve human wellbeing, while also respecting human rights*

2.10 Nonetheless, complexities can quickly arise when translating such principles into practice. There may be situations, for example, where an

22 Law Reform Committee – Subcommittee on AI and Robotics, *Report on the Application of Criminal Law to the Operation of Artificial Intelligence Systems and Technologies* (Co-chairs: Justice Kannan Ramesh & Charles Lim Aeng Cheng) (forthcoming).

23 Law Reform Committee – Subcommittee on AI and Robotics, *Report on the Attribution of Civil Liability for Accidents Involving Autonomous Cars* (Co-chairs: Justice Kannan Ramesh & Charles Lim Aeng Cheng) (forthcoming).

24 Draft Framework of Ethical Aspects, above n 8, at 6.

25 Jacques Berleur and Klaus Brunnstein (eds), Ethics of Computing: Codes, Spaces for Discussion and Law (Cham, Switzerland: Springer International Publishing, 1996) at 99. See also the Irish Computer Society Code of Professional Conduct (1994): Berleur and Brunnstein, *Id* at 120; the Computer Society of South Africa (CSSA) Code of Conduct: *Id* at 155; and the European Informatics Skills Structure (EISS – CEPIS) Code of Professional Conduct: *Id* at 175.

26 ACM Code of Ethics and Professional Conduct (New York, NY: Committee on Professional Ethics, Association for Computing Machinery, 22 June 2018), <*https://ethics.acm.org*> (accessed 10 June 2020).

27 Civil Law Rules on Robotics, above, n 8 at 22.

AI system would have to perform an act which is unlawful in order to avert causing injury to human beings. For instance, an autonomous vehicle may have to drive onto an empty pedestrian walkway to avoid colliding with a person.[28] Nor is there necessarily consensus (particularly between different cultures) as to how to resolve certain ethical questions – a case in point being the infamous 'trolley problem' of whether one life should be sacrificed to save five lives.[29] Accordingly, - and as will be discussed further in Section 0 below – policymakers will need to consider whether norms should be prescribed regarding how such scenarios should be resolved, and whether these should reflect any applicable international standards or practices or instead be culture-specific.

B. CONSIDERING EFFECTS

Designers and deployers of AI systems should consider the likely effects of AI systems throughout their lifecycle.

2.11 In common with the designers of non-AI products, those designing AI systems and their software life cycles should identify, analyse and address safety, security, legal, ethical and other issues arising from the intended and unintended effects of the AI system on people,[30] taking into account the complex environments within which the AI system will be deployed.

2.12 Similarly, it is accepted (whether or not viewed through the lens of 'ethics') that persons or organisations which intend to deploy AI systems should also conduct risk and impact assessments, evaluating the potential impact on affected stakeholders.[31]

---

28 See, for example, The Law Commission of England and Wales and the Scottish Law Commission (the "**UK Law Commissions**"), *Automated Vehicles: Summary of the Preliminary Consultation Paper* at [9.6]–[9.12], <*https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2018/11/6.5066_LC_AV_Final-summary_061118_WEB.pdf*> (accessed 10 June 2020).

29 Karen Hao, "Should a self-driving car kill the baby or the grandma? Depends on where you're from", MIT Technology Review (October 24, 2018): <*https://www.technologyreview.com/2018/10/24/139313/a-global-ethics-study-aims-to-help-ai-solve-the-self-driving-trolley-problem*>. Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon & Iyad Rahwan, "The Moral Machine experiment", Nature, (24 October 2018): <*https://www.nature.com/articles/s41586-018-0637-6*> (accessed 10 June 2020).

30 Compare the IEEE P7000 – Model Process for Addressing Ethical Concerns During System Design discussed in John C. Havens, Ethically Aligned Standards – A Model For The Future (13 March 2017) <*https://www.standardsuniversity.org/e-magazine/march-2017/ethically-aligned-standards-a-model-for-the-future/*>; EU High-Level Expert Group on Artificial Intelligence, "Ethics Guidelines for Trustworthy AI" (8 April 2019) ("**Ethics Guidelines for Trustworthy AI**") at 7, 17 <*https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai*> (accessed 10 June 2020).

31 See, for example, Model AI Governance Framework, above n. 1 at [3.7], which translates ethical principles into a governance and risk framework; Draft Framework of Ethical Aspects, above n 8, at 6.

2.13 An AI system should be rational, fair, and not contain biases that are intentionally or unintentionally built into their system which may harm a community of people or an individual.[32] Such irrationality, unfairness or bias could exist in the training data, the algorithm design, the design of the AI engine and/or the selection of a particular model for deployment.[33] Means of guarding against those risks may include conducting examinations of the AI system's input and output to determine and revaluate biases in their decision-making system. For example, a government agency which intends to use an AI system to assess a citizen's risk of committing certain types of offences should evaluate potential impact on fairness, justice, bias, and negative perceptions across affected communities, especially minorities.[34] A case in point is the controversial Northpointe AI system for risk assessment used in some US courts to aid judges in determining sentence. Evaluations of the software found that it wrongly predicted recidivism by black defendants at almost twice the rate at which it erred in relation to white defendants.[35]

2.14 Where an AI system produces wrongful effects, it will often (most likely, predominantly) be the case that the designer or deployer of the AI system did not intend for that effect to occur.[36] Only a small majority of those who design or deploy AI systems will do so with malicious intent. There could, for example, be pre-existing, undiscovered biases in the data on which the system was trained or on which it bases its decisions.[37] Indeed, given the complexity and 'opacity' of some AI systems' decision-making processes, the designer or deployer may not fully understand how the AI system arrived at a decision which produced the harmful effect. As

32 Compare the FEAT Principles, above, n 8 at 7-8, [5.3] and [5.6-5.7]; see Ethics Guidelines for Trustworthy AI, above, n 30 at 18; Draft Framework of Ethical Aspects, above, n 8 at 6.

33 Amazon's AI resume screening algorithm was found to be biased against women candidates because it was trained on data of applicants' resumes submitted to the company over 10 years which was apparently skewed towards a majority of men: Jeffrey Dastin, "Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women", Reuters, 10 October 2018 <*https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G*> (accessed 10 June 2020).

34 See Dillon Reisman, Jason Schultz, Kate Crawford and Meredith Whittaker, Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability (New York, NY: AI Now Institute, New York University, April 2018) at 4 <*https://ainowinstitute.org/aiareport2018.pdf*> (accessed 10 June 2020).

35 Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, "Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks.", ProPublica (23 May 2016) <*https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing*> (accessed 10 June 2020); Algorithms in the Criminal Justice System, Electronic Privacy Information Center <*https://epic.org/algorithmic-transparency/crim-justice/*> (accessed 10 June 2020).

36 For the avoidance of doubt, in this paper, terms such as designer, deployer, and user are not intended to be limited to individuals or any specific form of legal person.

37 See further paragraph 0 below.

noted above, however, this shifts the emphasis to questions of whether such effects were reasonably foreseeable, and – if they were – then:

(a) how that should impact the liability in law of the designer or deployer,[38] including the apportionment of such liability;[39] and

(b) the expectations on them in terms of the necessary and proportionate steps they must take to address it.[40]

2.15 It is possible that existing common law principles – for example (criminal or civil) negligence or other liability/equitable doctrines – are sufficient, and could be relied upon to fairly apportion liability in the scenarios described above. Fundamentally, however, the assessment for policy makers will be a broader risk management question, which may require more bespoke interventions setting out principles specific to certain scenarios (see, e.g. the Misrepresentation Act (Cap. 390) or the Frustrated Contracts Act (Cap. 115), for examples of equivalent interventions aimed at achieving equitable outcomes in specific scenarios).

C WELLBEING AND SAFETY

AI systems should be designed and deployed with the likely effects of AI systems, intended and unintended, assessed against holistic wellbeing and safety metrics.[41]

2.16 This principle of wellbeing and safety concerns the quality of life of all people, and the expectation that the design and deployment of AI should be for the benefit of society, its members, and the environment surrounding them. The corollary of this is that negative consequences of such systems,

38 Compare UKHL 2018 AI Report, above, n 13 at [309].

39 Related liability issues will be considered further in forthcoming reports in this Series, considering respectively, criminal liability and civil liability for accidents involving autonomous vehicles (above, n 22 and 23).

40 In the civil law context, the majority of the Singapore Court of Appeal in *Quoine Pte Ltd v B2C2 Ltd* [2020] SGCA(I) 2 observed that parties could enter into valid contracts through the application of deterministic algorithms (that is, algorithms that, given a particular input, will always produce the same output via the same steps). In such a scenario, the Court opined at [98] that the relevant state of knowledge for determining whether there was a mistake which may vitiate the contract is that of the programmer at the time of programming up to the point of contract formation. The broad point about whose knowledge and when could be extended to non-deterministic algorithms or systems. In the case of such systems, the designer or deployer ought to consider, given knowledge of all relevant circumstances at the material time, appropriate design or deployment parameters, human oversight or intervention within the system to ensure that potential wrongful effects would be avoided or controlled. See also *Model AI Governance Framework*, above n. 1 at [3.14].

41 See IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2 (Piscataway, NJ: IEEE, 2017) ("**Ethically Aligned Design, Version 2**") at 24–25; Civil Law Rules on Robotics at 22.

including threats to health, safety, personal security and privacy, should be minimised.

2.17 Generally, AI systems should at the very least 'do no harm' or minimise harm in unavoidable circumstances.[42] Negative effects on the same should be pre-empted and sufficiently mitigated.

2.18 Such principles have been articulated in existing ethical codes. As noted above, the ACM Code of Ethics and Professional Conduct (2018), for example, states that computing professionals should "contribute to society and to human well-being, acknowledging that all people are stakeholders in computing."[43] And the EU Committee on Legal Affairs Report similarly proposed that robot designers and engineers should "consider and respect people's physical wellbeing, safety, health and rights", "preserve human wellbeing […] and disclose promptly factors that might endanger the public or the environment […] while also respecting human rights."[44] Moreover, it proposed that engineers should not deploy a robot "without safeguarding the safety, efficacy and reversibility of the operation of the system." [45]

2.19 Policy challenges may arise from the fact that certain harms and negative effects are insidious and not capable of being well-defined. This is especially true of affective AI systems (also called "artificial emotional intelligence" or "emotion AI"), which aim to recognise, interpret, process, and simulate human emotions. These systems typically aim to simulate human empathy. Examples might include semi-humanoid robot nurses that provide company for isolated elderly people[46], or a call centre chatbot that uses voice analytics to gauge a person's emotions.[47] Affective AI systems therefore entail greater risks of causing unintended psychological effects and dependencies on people.[48] The difficulty for policymakers, however, is that such forms of harm do not lend themselves to simple codification in laws or regulation.

---

42 Discussion Paper on Artificial Intelligence (AI) and Personal Data – Fostering Responsible Development and Adoption of AI (Singapore: Personal Data Protection Commission, 5 June 2018) at 6 <*https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/Discussion-Paper-on-AI-and-PD--050618.pdf*> (accessed 10 June 2020); Model AI Governance Framework, above, n 1 at 65; Ethics Guidelines for Trustworthy AI, above, n 30 at 12; Draft AI R&D Guidelines, above, n 10 at [4.4] and [5.4].

43 ACM Code of Ethics and Professional Conduct, above, n 26.

44 Civil Law Rules on Robotics, above, n 8 at 22.

45 Civil Law Rules on Robotics, above, n 8 at 25.

46 See Sapiens, "The Rise of Emotional Robots", Sapiens.org (28 August 2018) <*https://www.sapiens.org/technology/emotional-intelligence-robots/*> (accessed 10 June 2020).

47 Meredith Somers, "Emotion AI, explained", MIT Management Sloan School (8 March 2019) <*https://mitsloan.mit.edu/ideas-made-to-matter/emotion-ai-explained*> (accessed 10 June 2020).

48 Civil Law Rules on Robotics, above, n 8 at 29.

2.20 AI systems should also be designed, as far as reasonably practicable, to avoid impinging on users' or third parties' privacy.[49] Privacy includes bodily integrity, physical or spatial privacy, privacy of personal data, and privacy of communications. If it is necessary to intrude on personal privacy, notice should be given or consent should be obtained.

2.21 In this respect, many jurisdictions, including Singapore, have existing data protection[50] or privacy laws which apply generally. Such rules play an important role in providing safeguards for individual rights and prompting those who develop AI systems to ensure privacy is 'designed in' to products and services, while maintaining sufficient latitude to enable innovation. Such laws will need to be kept under review, as technologies develop and cultural attitudes evolve, to ensure that they continue to provide effective protection of an individual's right to privacy (and thus of their wellbeing and safety).

2.22 While beyond the scope of this report, this Subcommittee has, in another report in this Series, separately considered whether a right of ownership or other enhanced rights should be granted over personal and/or non-personal data.[51]

2.23 Apart from the direct effects of AI systems, the indirect effects on human wellbeing may also need to be considered. These would include, for example, the indirect effect of AI systems on marginalised or disadvantaged communities within a society and whether any existing disadvantage or marginalisation would increase with the deployment of the AI system.[52]

2.24 Such insidious, indirect harms and marginalising effects may not be sufficiently severe or well-defined (particularly in a rapidly evolving landscape) to be addressed by way of criminal legislation. Hence, alternative forms of domain-specific intervention may need to be considered to address them.

2.25 As AI systems become more prevalent and more complex, increased consideration will also likely need to be given to the development of national and international standards for AI systems.[53] Standards can serve to ensure consistent protection of individuals' wellbeing and safety across

---

49 Draft AI R&D Guidelines, above, n 10 at [4.6] and [5.6]; Draft Framework of Ethical Aspects, above n 8, at 6.

50 In Singapore, the Personal Data Protection Act 2012 ("PDPA"), and in the EU, the General Data Protection Regulation ("GDPR"). Consent is the touchstone of such legislation: see e.g. section 13, PDPA; article 7, GDPR.

51 Law Reform Committee – Subcommittee on AI and Robotics, *Rethinking Database Rights and Data Ownership in an AI World* (July 2020) (Co-chairs: Justice Kannan Ramesh & Charles Lim Aeng Cheng) <*https://www.sal.org.sg/Resources-Tools/Law-Reform/Law-Reform-e-Archive*>.

52 Compare the FEAT Principles, above, n 6 at [5.1]; Draft Framework of Ethical Aspects, above n 8, at 6.

53 Draft Framework of Ethical Aspects, above n 8, at 8.

borders, and avoid a 'race to the bottom' on standards and/or risks of regulatory 'arbitrage' by those that design or deploy AI systems. Standards may also help to promote interconnectivity and interoperability of AI systems,[54] leading to better co-ordination between such systems and thus a lower risk of conflicts, accidents and harm. Further, interoperability and interconnectivity can facilitate service and data portability, which should, in turn, help stimulate adoption of beneficial AI systems.[55]

2.26 Examples of interconnectivity and interoperability include the standardisation of data formats, the adoption of universal protocols for AI systems to communicate, and non-personal data sharing among AI systems. Governments can often help drive or incentivise such standardisation, as was the case with the adoption of standardised QR codes for digital payments in Singapore.[56]

D RISK MANAGEMENT

It is imperative for designers and deployers of AI systems to properly assess and eliminate or control risks of the use of AI systems as a matter of safety and wellbeing.

2.27 Inherent in this principle is the proposition that AI systems and robots should undergo rigorous testing (proportionate to the extent of risks their deployment entails) in: (a) laboratory test beds that let users interact with and evaluate them; and (b) controlled real-world environments to ensure optimal controllability.[57] This will uncover risks relating to an AI system and also assist the AI system designer and deployer to mitigate and reduce the magnitude of harm if an accident does occur during or after deployment.

2.28 Policymakers will need to consider whether mandatory risk management standards need to be imposed, and if so, the form in which, and specificity with which, such standards are articulated. This may include, for example, mandating or promoting the use of certain risk mitigation mechanisms for specific types of AI systems, such as the installation of fail-safe components (action limitations and tripwires).[58] However, where AI systems carry particularly significant or unpredictable

54 Draft AI R&D Guidelines, above, n 10 at [4.1] and [5.1]; Ethics Guidelines for Trustworthy AI, above, n 30 at 11.

55 See further Law Reform Committee – Subcommittee on AI and Robotics, *Rethinking Database Rights and Data Ownership in an AI World* (July 2020) (Co-chairs: Justice Kannan Ramesh & Charles Lim Aeng Cheng) at [2.55] <*https://www.sal.org.sg/Resources-Tools/Law-Reform/Law-Reform-e-Archive*>.

56 See MAS website on Singapore Quick Response Code (SGQR): <*https://www.mas.gov.sg/development/e-payments/sgqr*> (accessed 10 June 2020).

57 Ethically Aligned Design, Version 2, above, n 41 at 50; Draft AI R&D Guidelines, above, n 10 at [4.3] and [5.3].

58 Ethically Aligned Design, Version 2, above, n 41 at 47 and 79.

risks of negative effects, policymakers may need to consider whether such systems should be required to be 'safe by design' so as to prevent the occurrence of such risks, rather than rely only on fail-safe mechanisms that operate reactively when the risks occur.

2.29 Workplace safety legislation provides an example of such 'safe by design' requirements in action. The Workplace Safety and Health (Design For Safety) Regulations 2015 (Cap 354A), for example, impose obligations on:

(a) designers to "as far as it is reasonably practicable, prepare a design plan for the structure that eliminates all foreseeable design risks;[59] and

(b) project developers to "as far as it is reasonably practicable, ensure that all foreseeable design risks in the project are eliminated".[60]

Further, where it is not reasonably practicable to eliminate a foreseeable design risk, project developers (or, as the case may be, designers) must – broadly stated – take steps to reduce the design risk (at source, and using collective, rather than individual, protective measures[61]) to as low as is reasonably practicable.[62]

2.30 The extent to which risk assessments and mitigations should be evaluated and documented may also need to be reconsidered, so as to enable effective ex-post investigation and 'auditability' (see further Section 0 below) should harm occur. Again, workplace safety legislation provides one example of how such actions are presently mandated, imposing as it does duties on project developers to conduct risk assessments, eliminate and control risks, and keep records of the same.[63]

2.31 In an AI context, risks that may need to be evaluated, documented, and mitigated include, for example:[64]

(a) ***Unintended side effects***. For example, if an AI system is designed to achieve a specific objective within a system –

59 Workplace Safety and Health (Design For Safety) Regulations 2015 (Cap 354A), Regulation 9. "Design risks", in this context (that is, relating to structures) are anything present or absent in the design of the structure that increases the likelihood that an affected person will suffer bodily injury when constructing, working at or demolishing the structure" (Regulation 2).

60 *Id*. Regulation 4.

61 Collective protective measures are those that protect everyone or a group of people. Individual protective measures protect an individual, and typically require that person to take some action to protect themselves.

62 *Id*. Regulations 4(2) and 9(2).

63 *Id*. Regulations 4-7

64 From Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman and Dan Mané, Concrete Problems in AI Safety (Ithaca, NY: arXiv, Cornell University, 25 July 2016) <*https://arxiv.org/abs/1606.06565*> (accessed 10 June 2020).

such as a robot designed to move a mug from the dining room to the kitchen – there could be many different obstacles between the robot and the fulfilment of its objective. In pursuing its objective, will the robot be able to identify a safe way to circumvent the obstacles? What if the robot has to roll over and crush an object in order to achieve its objective?

As it would be virtually impossible to foresee *all* possible obstacles which may arise between the robot and its objective, the robot must be designed in such a way that it can deploy general approaches to address potential obstacles.

Reward hacking. That is, where an AI system that is designed to achieve a certain objective 'hacks' the means to achieve that objective in an unintended way. For example, a robot that is 'rewarded' (that is, programmed to achieve specific goals) to clean dirt may continually dump the dirt it has just cleaned in a small area of a house so that it can pick it up again and continue to obtain the reward.

Again, insofar as it is difficult to foresee all the possible ways an AI system may hack its reward system, it is important for AI system designers to consider appropriate strategies to pre-empt such behaviour.

(b) ***Scalable oversight***: Where an AI system is designed to assess its performance in achieving an objective using proxy feedback, problems may arise if that proxy information does not, in fact, accurately reflect the system's actual performance. That disconnect may cause the system to act in unintended ways and/or cause particular unintended risks (aggravated or otherwise) to manifest.

Take, for example, a robot designed to clean the house in a way "that its user will be pleased with". If the user does not provide sufficient feedback about the acceptability of its cleaning, the robot will have to rely instead on proxy indicators, such as the amount of dirt it sees in the house. However, relying on such imperfect proxies may in fact cause the robot to wrongly identify certain material in the furniture as dirt, and thus inadvertently to cause damage to that furniture.

Evidently, where such a scenario is translated from a relatively benign, domestic example to something much broader (for example the public deployment of AI systems at scale across a community), the potential negative effects are much greater than mere damage to furniture.

(c) ***Safe exploration*** – AI systems that undertake exploration[65] will typically be designed to undertake such exploration safely. However, it is impossible to foresee every possible danger in such exploration, and thus there remains a risk that AI systems deployed in the real world which undertake exploration may cause harm to people, other objects, or themselves.

(d) ***Robustness to distributional shift*** – problems may arise where an AI system is placed in an environment which differs significantly from the environment it was conceived and trained to operate in. For example, if a robot designed to clean a house was trained to clean only floors and furniture, it might do something destructive when it encounters a toy or a pet, and still assume that it is performing well.

2.32 As the foregoing demonstrates, there may be risks of unintended consequences, not all of which will be reasonably foreseeable by the designers or deployers of AI systems before the deployment of the system. As previously discussed, from a policymaking perspective, this engages questions regarding the standards to which designers and deployers should be held, the level of risk that a society is willing to accept, the possible attribution of civil and criminal liability (and the attendant challenges thereof), and how the costs of damage or harm caused by such unintended consequences should be distributed.

2.33 Again, various approaches to addressing those questions may be taken, including, among others:

(a) Mandating that certain types of organisations, e.g. those that design and deploy AI systems, take up product and accident liability insurance;[66]

---

65 Exploration, in this context, refers to the deployment of an AI system (for example a robot) into an unknown environment, with the goal of the robot, as it moves through that environment, acquiring information and reducing uncertainty about the environment. See Nathan Michael, 'How Artificial Intelligence Manifests Through Exploration', Shield AI (17 July 2019) <*https://www.shield.ai/content/2019/7/15/how-artificial-intelligence-manifests-through-exploration*> (accessed 10 June 2020).

66 For a useful discussion of the benefits and challenges of a mandatory insurance in an AI context, see *Liability for Artificial Intelligence and Other Emerging Digital Technologies*, Expert Group on Liability and New Technologies – New Technologies Formation (November 2019) ('***Liability for AI and New Technologies Report***') <*https://ec.europa.eu/transparency/regexpert/index.cfm?do=groupDetail.groupMeetingDoc&docid=36608*> (accessed 10 June 2020). See also: the United Kingdom's Automated and Electric Vehicles Act 2018 c 18 (UK), which, for autonomous vehicles, mandates insurance under a 'single insurance' policy covering both the driver and the vehicle; and the UK Law Commissions' joint 2018 consultation paper on automated vehicles, which proposes that the organisation putting an autonomous driving system forward for authorisation (typically being the vehicle manufacturer or the developer of the automated driving system) be required to demonstrate that it carries appropriate liability insurance. UK Law Commissions, *Automated Vehicles:*

(*cont'd on the next page*)

(b) Establishing a certification regime whereby AI systems must be certified before deployment and certifiers could be held liable for the effects of the AI systems;[67]

(c) Employing a fault-based system, equivalent to that embodied in existing common law tortious rules in Singapore; or

(d) Imposing safety obligations and prescribing minimum safety standards for the design and deployment of AI systems, similar to workplace safety standards.[68]

Each approach has advantages and disadvantages, and will be more suited to some scenarios or sectors than others. In particular areas, a combination of approaches may be necessary or beneficial to achieve an optimal outcome.[69]

E VALUES AND CULTURE

AI systems should be designed to take into account, as far as reasonably possible, societal values and the cultural context of the environment in which they will be deployed.[70]

2.34 Societal values and cultural norms may be embedded into AI systems as values by design or through inadvertent bias. For example, an AI system designer may design the system to respond to certain categories of personal data such as ethnicity, religion, or language in a certain manner, or may have omitted to design appropriate responses to such categories

---

A Joint Preliminary Consultation Paper (Law Commission Consultation Paper 240, 2018) (Chairman: Mr Justice Green) at [1.44] <*https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2018/11/6.5066_LC_AV-Consultation-Paper-5-November_061118_WEB-1.pdf*> (accessed 10 June 2020); Simon Chesterman, Artificial Intelligence and the Problem of Autonomy (April 7, 2020). Notre Dame Journal on Emerging Technologies, 1 (2020), 210–250; NUS Law Working Paper No. 2019/016 at 9-12. <*http://dx.doi.org/10.2139/ssrn.3450540*> (accessed 10 June 2020).

67 See Changing Driving Laws to Support Automated Vehicles: Policy Paper (Melbourne, Vic: National Transport Commission (Australia), 2018) at 37 <*https://www.ntc.gov.au/sites/default/files/assets/files/NTC%20Policy%20Paper%20-%20Changing%20driving%20laws%20to%20support%20automated%20vehicles.pdf*> (accessed 10 June 2020); Draft Framework of Ethical Aspects, above n 8, at 8.

68 See the Workplace Safety and Health Act (Cap. 354A)

69 For example, the European Commission recently sought views on, inter alia, whether AI applications "with a specific risk profile" should be subject to both strict liability and an obligation to be insured, so as to ensure compensation irrespective of the liable person's solvency and help reduce the costs of damage. European Commission, *Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics*. COM(2020) 64 (February 2020) at 16 <*https://ec.europa.eu/info/sites/info/files/report-safety-liability-artificial-intelligence-feb2020_en_1.pdf*> (accessed 10 June 2020).

70 See Ethically Aligned Design, Version 2, above, n 41 at 58–59, 164–171 and 203–211.

(potentially biasing outcomes).[71] Alternatively, it may be the case that an AI system is trained on data that was inherently biased or unrepresentative, and thus internalises such biases in all its decision making.[72]

2.35 A society's norms and cultures may also impact how it views, or balances, many of the ethical questions already raised in this report. For example, different societies and cultures have different perceptions of privacy, which may impact the requirements imposed on the design of AI systems (see further section C above) or – as noted in paragraph 0 above – they may provide diametrically opposed answers to ethical dilemmas such as the trolley problem. The same is true for issues such as how conflicts between different values or ethical principles should be resolved by AI systems in determining how to act, and the extent to which such resolution processes need to be explained to users.

2.36 Further, societal values and cultural norms may evolve within a particular culture or across cultures over time. Policymakers will therefore need to give thought not only to how any regulatory interventions can accommodate, or be adapted to account for, such changes, but also whether (and if so to what extent) AI systems themselves are required to be designed to be similarly capable of being updated in response to such changes.

2.37 Taking into account societal values and cultural norms is especially important in affective AI systems. Verbal and non-verbal communication to be made by an affective AI system should take into account cultural sensitivities. For example, certain slang or colloquial terms have special meaning in a specific culture. Norms regarding matters such as eye contact, level of personal space, facial expressions, physical gestures, physical greetings or appropriate types of physical touch may also vary across cultures. As will be evident, the potential for negative effects if inadequate safeguards are in place is significant.[73]

2.38 While not the only potential solutions, two possible options for a regulatory response to this issue are pre-deployment approval and *ex post facto* action. Unlike a film or play, an AI system cannot practicably be reviewed beforehand for culturally inappropriate effects or communications. A relevant authority may, therefore, require a designer or deployer of an AI system to self-disclose potential risks and issues in this respect and to conduct an audit of the system with reference to those self-

71 Ethics Guidelines for Trustworthy AI, above, n 30 at 18.

72 See further paragraph 0 above.

73 An infamous example is the Microsoft "Tay" AI chatbot. After Tay was deployed, it was exposed to racist, misogynistic, political and offensive comments from the public and began to make similar statements. 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 10 June 2020).

disclosed risks and issues. Alternatively, the authority may rely on *ex-post* enforcement action where harms arise after deployment (noting that, as previously highlighted, in seeking to do so, questions as to who should be held responsible for certain consequences of AI systems, and to what extent, will need to be addressed[74]).

2.39 Finally, it merits stating that, particularly where the deployment of an AI system requires value judgments which are wholly novel (or for which there is as yet no range of 'accepted' cultural approaches to particular ethical questions) opportunities may arise for national policymakers to use their policy interventions to set a 'high water mark'. This, in turn, could serve to encourage an ethical 'race to the top' in AI-related regulation internationally.

E TRANSPARENCY

AI systems should generally be designed to be transparent as far as reasonably possible.[75] *It should be possible to discover how and why an AI system made a particular decision or acted the way it did.*

Transparency also entails a need to ensure, as far as reasonably possible, traceability, explainability, verifiability and interpretability of AI systems and their outcomes.

2.40 Transparency is essential in order for AI systems to be properly regulated, audited, deployed and accepted by users, stakeholders and the wider public. Furthermore, increased transparency would more likely translate to increased public confidence in those systems and facilitate quicker and more widespread adoption.

2.41 Also, in the event of any dispute or potential legal liability, transparency is essential for relevant stakeholders, parties, and the courts or other adjudicating bodies, to resolve such disputes and liability issues. Transparency would also facilitate the work of fact-finding tribunals such as Coroner's Inquiries and Committees of Inquiry. More fundamentally still, transparency is integral to the rule of law, in particular where the AI system is used to facilitate the discharge of a public duty or responsibility.[76]

74 For example, if an AI system made a communication which would, if it were a human being, constitute an offence of deliberate intent to wound the religious or racial feelings of a person under section 298 of the Penal Code, should the designer and/or deployer of the AI system be held liable?

75 Ethically Aligned Design, Version 2, above, n 41 at 29–30 and 45–46; Discussion Paper on Artificial Intelligence (AI) and Personal Data, above, n 42 at 5; Draft AI R&D Guidelines, above, n 10 at [4.2] and [5.2]; Civil Law Rules on Robotics, above, n 88 at 10; Ethics Guidelines for Trustworthy AI, above, n 30 at 18; Model AI Governance Framework, above, n 1 at 15; FEAT Principles, above, n 6 at 12.

76 See further below, paragraph 0 and n 83 and n 90.

2.42 Looking at the individual elements of transparency, 'traceability' and 'explainability' requirements may, for example, entail an obligation to ensure as far as reasonably possible that the complex processes, actions or 'thinking' of AI systems:[77]

(a) are documented in a way that is understood easily, or

(b) can be explained when questioned using a standard human cognitive approach, such that the logic of those processes and decisions can be understood in non-technical terms.

Similarly, 'verifiability' requirements could involve duties to ensure that the output of AI systems can be verified as in accordance with its intended processes and parameters.

2.43 As will be apparent, therefore, 'opaque' decision-making systems, software services or components (also known as 'black boxes') may give rise to particular challenges in this regard. In such scenarios, additional measures may be needed to: (a) mitigate the risks involved (e.g. by setting strict and well-defined parameters for safe deployment);[78] or (b) impose proportionate obligations on those designing or deploying the systems.

2.44 Further, it is often said that AI systems technically face trade-offs: (i) between explainability and accuracy, and (ii) between predictability and accuracy.[79] This presents a challenge for regulation, insofar as it implies that indiscriminate universal approaches are likely to be inappropriate.

2.45 One possible regulatory response to traceability, explainability and verifiability challenges is to require mechanisms to be built into AI systems that, as far as reasonably possible, record input data and provide a logic behind decisions taken by the AI (acknowledging that for certain complex deep learning systems, full explainability may not be useful or achievable). A well-known analogue for this exists in the obligation to fit a flight data recorder in airplanes, and an increasing number of national regulators have imposed or are considering similar duties for the autonomous vehicles currently being introduced to public roads.[80] Indeed, an Expert Group of the European Commission recently similarly recommended that emerging

77 Draft Framework of Ethical Aspects, above n 8, at 5.

78 Ethically Aligned Design, Version 2, above n 41 at 70–71.

79 For example, the machine learning model adopted in an AI system may be rule-based and thus more explainable and predictable, but less able to take into account a wide range of variants in the data. See Hacker, P., Krestel, R., Grundmann, S. et al. Explainable AI under contract and tort law: legal incentives and technical challenges. Artif Intell Law (2020): <*https://link.springer.com/article/10.1007/s10506-020-09260-6?shared-article-renderer*> (accessed 10 June 2020).

80 For example, Germany (see §63A, Straßenverkehrsgesetz (StVG) (Germany)) and Japan (See Dan Matsuda, Edward Mears and Yuji Shimada 'Legalization of Self-Driving Vehicles in Japan: Progress Made, but Obstacles Remain', DLA Piper (18 June 2019) <*https://www.dlapiper.com/en/japan/insights/publications/2019/06/legalization-of-self-driving-vehicles-in-japan/*>) (accessed 10 June 2020)).

technologies (not just autonomous vehicles) should be required to have data recording capabilities (termed 'logging by design'), where a) such information is "typically essential for establishing whether a risk of the technology materialised," and b) requiring such recording is appropriate and proportionate. Furthermore, the group recommended that if such data is not recorded or made reasonably accessible, the burden of proof should be reversed in favour of the person harmed.[81]

2.46 Alternatively, policymakers might consider adopting mechanisms to ensure that AI systems (or those in specific sectors or with specific risk profiles) are audited and certified by specialised third-party agencies prior to launch.[82]

2.47 Transparency also implies that AI systems should be (in practice, and by design) 'honest' in their interaction with users. This means that an AI system must correspond to the functions it was created to do and its functions as communicated to others. Emphasis therefore then has to be placed on the sufficiency and accuracy of the information that AI system designers and/or deploying organisations convey to end-users, to each other and to other stakeholders about the AI system. In this regard, it must be borne in mind that AI system designers and deployers may be separate entities and that the designers may not, for example, have control over either the data used for "deep learning" by the AI system and/or the dataset on which the AI system operates.

2.48 In certain contexts, the need for a high level of transparency of AI systems is much greater, e.g. judicial affairs, public administration, healthcare, personal loans and insurance, autonomous vehicles, public transport systems and weapon systems.[83] While the general principle of transparency in AI systems is widely accepted, there are differing views regarding the feasibility and necessity of mandating such transparency in particular contexts, as well as how it should be evidenced (including, for example, the extent to which organisations should be required to disclose the fact that AI systems are used in making decisions or that the individual is interacting directly with an AI system).[84] Different contexts may require different levels of transparency and different approaches to ensuring it. For example:

- In certain contexts, informational requirements and pre-approval mechanisms may be effective. In the non-AI domain, such mechanisms currently exist in sectors such as

81 See Liability for AI and New Technologies Report, above n 66 at 47-48.
82 Ethically Aligned Design, Version 2, above, n 41 at 160.
83 UKHL 2018 AI Report, above, n 13 at [94]; see also Makoto Hong Cheng and Hui Choon Kuen, "Towards a digital government: Reflections on automated decision-making and the principles of administrative justice" [2019] SAcLJ, Vol. 31, No. 2, 875.
84 UKHL 2018 AI Report, above, n 13 at [95]-[99].

pharmaceutical licensing and labelling, and the sale of consumer financial products. [85]

- In relation to automated decision making, the United Kingdom Information Commissioner's Office considers it best practice, in line with the EU General Data Protection Regulation ('**GDPR**'), for companies to tell customers about the profiling and automated decision-making they carry out, and the nature and source of the information used to create profiles regarding individuals.[86]
- In California, laws have been introduced to "make it unlawful for any person to use a bot to communicate or interact with another person in California online with the intent to mislead the other person about its artificial identity for the purpose of knowingly deceiving the person about the content of the communication in order to incentivise a purchase or sale of goods or services in a commercial transaction or to influence a vote in an election".[87]

2.49 Such transparency requirements are easily stated, but present difficulties when applied in practice. For example, challenges may arise in defining precisely what situations should be disclosed, given that the type and extent of use of AI systems vary widely, as do their levels of automation in making or implementing decisions.[88] As a case in point, the Californian laws have been criticised on various grounds, including the scope of bots that are covered, the inherently unpredictable nature of communications made by certain types of automated messaging bots (which makes it difficult to determine, for example, whether such communication counts as

85 See, for example, Medicines (Labelling) Regulations (Rg 5, S 255/1986. 2000 Rev Ed), and Consumer Protection (Fair Trading) (Regulated Financial Products And Services) Regulations 2009 (CAP 52a, S 54/2009).

86 See Information Commissioner's Office, *Guide to the General Data Protection Regulation* (May 2019) at 160 <*https://ico.org.uk/media/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr-1-0.pdf*> (accessed 10 June 2020). Under Article 22(1) of the EU General Data Protection Regulation (GDPR), individuals have the right (subject to some exceptions) not to be subject, without explicit consent, to a "decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly affects him or her."

87 Senate Bill No. 1001, Chapter 892: <*https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001*> (accessed 10 June 2020).

88 See Jorgen Frohm, Veronica Lindstrom, Mats Winroth, Johan Stahre, "Levels of Automation in Manufacturing" International Journal of Ergonomics and Human Factors (2008), Vol. 30, Issue 3 at 10-12, which discusses, among others, a ten-level taxonomy of automation by Sheridan and Verplanck that focuses on whether it is the human or the machine that generates alternatives, makes decisions and implements decisions.

'commercial' or 'political'), and even concerns about the chilling of free speech.[89]

2.50 Furthermore, 'technical' transparency may not actually serve to protect an individual's right – a complex explanation of the specific decision-making processes undertaken by an AI system may well mean little to an ordinary, non-expert person. In that regard, it is noted that the GDPR (Art 22) and its UK analogue, the Data Protection Act 2018 (s.49-50) require that companies (subject to narrow exceptions) enable individuals both to challenge an automated decision that has legal effects on him/her, and to 'opt out' from such automation and request human intervention.[90] In the Singapore context, the guidance in the FEAT Principles[91] and the Model AI Governance Framework[92] suggest giving customers explanations of decisions of AI systems and processes and also the opportunity to request a review of such decisions.[93] It may be that such rights are both more practically feasible and more helpful than technical transparency and better help to ensure fairness in how AI systems impact individuals (see further paragraph 0 above).

G ACCOUNTABILITY

Those who design and deploy AI systems should be accountable for the proper functioning of those systems and for their respect of the ethical principles detailed herein, based on their roles, the context, and consistency with the state of art.[94]

2.51 Policymakers will need to consider the best mechanism for ensuring due accountability and for facilitating recourse and just apportionment of liability between AI system designers, manufacturers, deployers and users when wrongs happen.

2.52 As previously alluded to, given the complex way in which numerous actors' contributions to an AI system might interrelate and impact one another, establishing which aspects of the AI's design or operation contributed to a wrong, to what extent and who is responsible for them,

89 Dave Gershgorn, "A California law now means chatbots have to disclose they're not human" (4 October 2018): <*https://qz.com/1409350/a-new-law-means-californias-bots-have-to-disclose-theyre-not-human/*> (accessed 10 June 2020).

90 Although not considered in detail here, such questions of transparency and automated decision making may be of particular significance in relation to public administrative decisions (for example, decisions on visa applications) made without any human involvement. See further Makoto Hong Cheng & Hui Choon Kuen, "Towards a Digital Government: Reflections on Automated Decision-making and the Principles of Administrative Justice", above, n 83.

91 FEAT Principles, above, n 6 at [8.2].

92 Model AI Governance Framework, above, n 1 at [3.48]-[3.50].

93 *Id* at [3.53]; FEAT Principles, above, n 6 at 11.

94 Model AI Governance Framework, above, n 1 at Annex A.

might be factually – let alone legally – extremely challenging. For instance, AI system designers, system integrators, deployers and users may have different roles and degrees of contribution to the design, testing, training, calibration and use of an AI system.

2.53 In circumstances where AI systems are deployed in the context of clearly-defined relationships, e.g. the provision of banking services to a customer, it may be more apparent where responsibility lies and there may be 'traditional' avenues in contract or other laws through which customers may seek recourse. However, the same may not be true where those AI systems are deployed in a manner that exposes them to the wider public.

2.54 Similar accountability challenges have previously been addressed by (among other mechanisms) registration and records schemes – a prominent example being vehicle registration systems. It is notable that one of the justifications for mandatory registration of e-scooters in Singapore's recent Active Mobility Act[95] was that it would "help deter reckless behaviour, accord more responsibility to the users, and facilitate enforcement officers in tracking down errant users".[96] In principle. a similar objective could be pursued by a scheme requiring the key identifiers and parameters of certain AI systems (for example those entailing particularly high risks of causing harm) to be registered and recorded, so as to enable those affected by the system to identify and trace it, and the parties responsible for it.

H DATA

AI system designers and deployers should respect laws that protect the personal data of individuals and observe good data practices to ensure that data is collected accurately, stored safely, processed fairly, and disclosed only where necessary. Proper records of data provenance should be kept.[97]

2.55 In many jurisdictions, including Singapore, data protection legislation would already apply to personal data which may be collected or processed in relation to the development and deployment of AI systems. Nevertheless, constant review and finetuning of such legislation – as well as consideration of other 'soft' regulatory mechanisms – will likely be required to ensure that AI system designers and deployers have the regulatory freedom to research, develop and deploy AI systems but remain subject to sufficiently robust safeguards to effectively protect the general public.[98]

95 Active Mobility Act 2017 (Act 3 of 2017).

96 Senior Minister of State for Transport (Dr Lam Pin Min), Committee of Supply - Head W (Ministry of Transport) (7 March 2018), Parl. 13, Session No. 1, Vol. 94, Sitting 67.

97 Model AI Governance Framework, above, n 1 at 17, 35-40.

98 See e.g. the Personal Data Protection Commission's ("PDPC") intended amendment of the PDPA to introduce data innovation provisions, which will a) provide for certain "derived data" to be excluded from the Access, Correction and proposed Data

(*cont'd on the next page*)

CHAPTER 3

CONCLUSION

3.1 The diverse and constantly evolving development of AI technology and systems and the ecosystem in which they operate require the right balance to be struck between fostering innovation and growth on the one hand, and building public trust by way of managing risks on the other, such that all segments of society are able to harness digital technologies and benefit from them.

3.2 Given that the goal of widespread deployment of AI systems is ultimately about improving people's lives, it is right – and increasingly accepted – that such deployment should be human-centred and built on strong ethical foundations. That in turn has implications for the laws and other regulatory interventions which enable, guide and constrain such deployment.

3.3 The various considerations and principles discussed in this report do not seek to advance any specific means or level of intervention. Indeed, it is likely that different regulatory approaches will be required in relation to different technologies, sectors or contexts. However, it is hoped that this report's analysis provides a framework for a public discussion on the best means to achieve human-centred, ethical norm-making and calibration of regulatory responses in respect of AI.

3.4 We recognise that regulatory intervention may not reach the many AI systems developed overseas and used in Singapore. It is hoped, however, that an international consensus on principles for a common framework will emerge and that that will, in turn, inform the conduct of designers of AI systems developed overseas. By developing our own framework, Singapore would have contributed in a small way to developing this international consensus.

Portability provisions and b) clarify when organisations may use personal data without consent for appropriate business improvement purposes. PDPC, "Response To Feedback On The Public Consultation On Proposed Data Portability And Data Innovation Provisions" (20 January 2020) at 14-15 <*https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Legislation-and-Guidelines/Response-to-Feedback-for-3rd-Public-Consultation-on-Data-Portability-Innovation-200120.pdf*> (accessed 10 June 2020).

GLOSSARY[99]

Affective AI (also sometimes referred to as **Artificial Emotional Intelligence**, **Emotion AI** or **Empathic AI**) – **AI systems** that aim to recognise, interpret, process, and simulate human emotions, in particular human empathy.

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).[100]

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).

99 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.

100 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 1.

Big Data — datasets characterised by their:

a) size ("Volume");

b) complexity ("Variety") (i.e. typically including structured, semi-structured and unstructured data derived from diverse sources); and/or,

c) 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.

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.

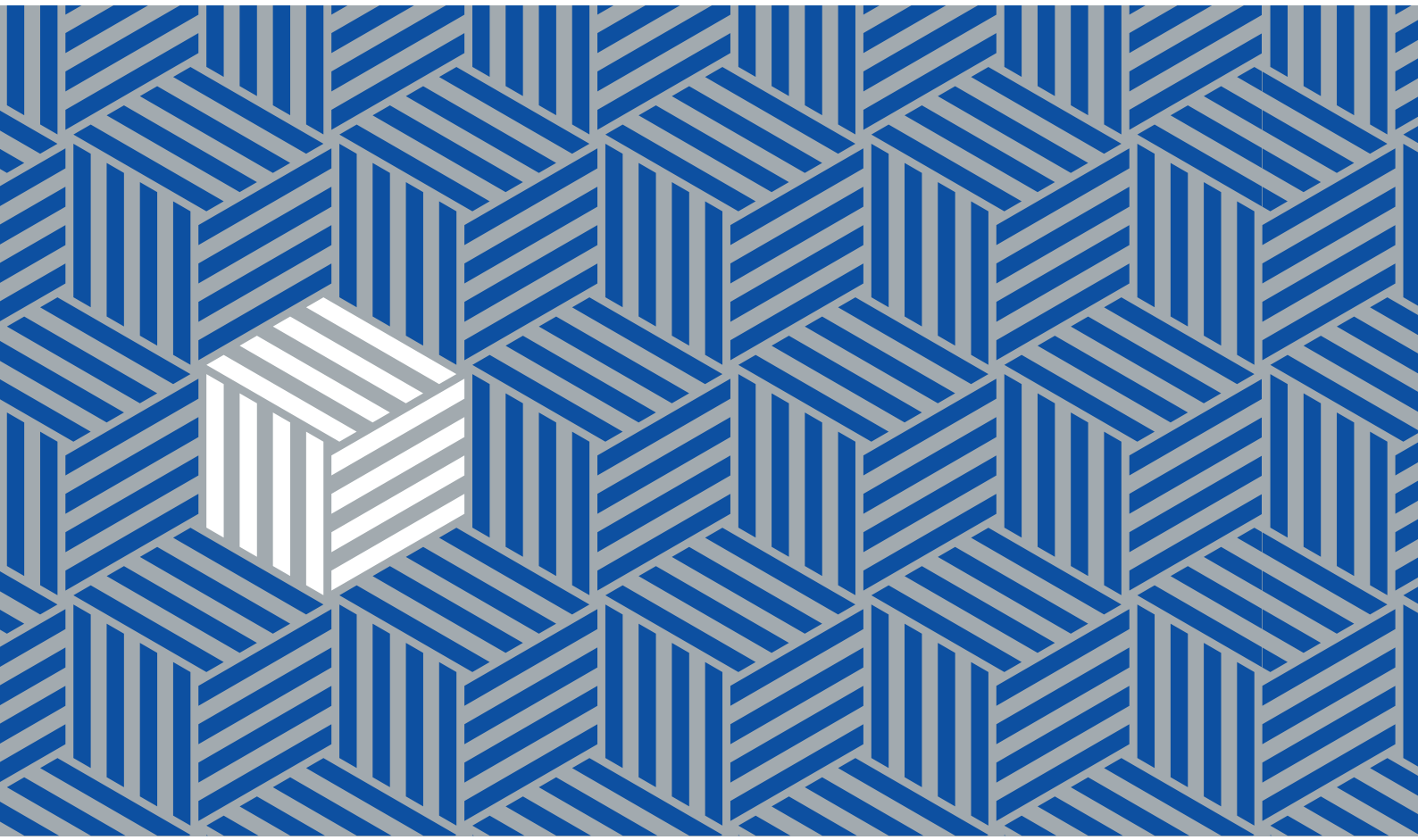
Transparency — various mechanisms or requirements intended to provide additional information to users, regulators and other stakeholders regarding the algorithmic decision-making processes undertaken by **AI systems**, and the input **data** relied on by such systems. Such transparency may be achieved 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**).

Trolley Problem, The — an ethical dilemma which – in its most common form – asks whether a person should reroute a runaway trolley, causing it to hit and kill one person, if not rerouting the trolley would have meant it hit and killed five people.

Turing Test, The — a test which tests a machine's ability to exhibit behaviour that is indistinguishable from that of a human. An **AI system** would pass the Turing Test if the human tester could not reliably distinguish the **AI system**'s responses to natural language questions from those of a human.

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 978-981-14-6600-7 (softcover)
978-981-14-6601-4 (e-book)