A 'Human-Centric and Lifecycle Approach' to Legal Responsibility for AI

*Ozlem Ulgen**

Introduction

Artificial intelligence (AI), involving use of algorithms and robotics to perform tasks that would otherwise require human intelligence, presents challenges as to who should be held legally responsible for any resulting loss, damage, harm, or injury. Lacking legal agency yet having causal agency, AI may disrupt the causation chain and attribution of responsibility unless there is always a human who will be held responsible throughout AI design, development, and deployment stages. The humanly inexplicable and unattributable behaviour of self-learning algorithms, robots, and 'black box' scenarios amplifies harm potential. AI without human legal responsibility undermines human agency of foresight, prudence, and judgement in taking action with consequences in mind. This article develops a 'human-centric and lifecycle approach' to establishing legal responsibility for AI, based on the role of human agents and the user-system lifecycle. Part 1 explains the problem of unintended consequences in AI, and the functionality and temporality of legal responsibility to reflect sanctions, victim interests, and wider societal values. It is argued that functionality and temporality enable locating past and future responsibility of various human agents at the pre-deployment and post deployment stages. Part 2 evaluates a combination of precautionary and strict liability models for allocation of responsibility and liability in the field of AI. Given AI's lack of legal agency yet potential to cause adverse effects reaching the standard of loss, damage, harm or injury, consideration is given to whether it is necessary to establish causation or adopt a prospective approach to obligations and sanctions regardless of fault. Part 3 develops the content of a 'human-centric and lifecycle approach' to responsibility based on prioritisation of human interests, user awareness, protection of rights, respecting AI limitations, and non-deterministic influences on decision-making. Under this approach, responsibility is represented throughout the system lifecycle (e.g. design; development; and deployment) and attributable to a broad spectrum of human agents (e.g. designers; programmers; engineers; manufacturers; operators; and system owners).

1. Unintended consequences and the function of responsibility

In 2018, whilst working as a technical expert to develop international standards for use of AI, I came across a software programming ontology, WordNet Synsets, which referred to the concept of obligation through the hyponym 'the white man's burden'.[1] Rudyard Kipling coined the phrase in his 1899 poem, 'The White Man's Burden: The United States and the Philippine Islands', encouraging the United States to assume colonial control over the Filipino people and their country. The phrase is associated with racial prejudice and colonial exploitation, corresponding with Kipling's belief that the British Empire was the Englishman's 'Divine Burden to reign God's Empire on Earth'.[2] The WordNet Synsets classification illustrates the potential for programming bias to creep into systems design and development through use of outdated, culturally insensitive, and

* Reader in International Law and Ethics, School of Law, Birmingham City University, UK. ozlem.ulgen@bcu.ac.uk. A version of this paper was presented at the ILPC Conference, London, 19-20 November 2020. The author would like to thank the discussant and participants for their critical feedback.

[1] WordNet Synsets <101129920, http://sigma.ontologyportal.org:8080/sigma/WordNet.jsp?synset=101129920&kb=SUMO&flang=SUO-KIF&lang=EnglishLanguage&kb=SUMO> accessed 17 February 2021.

[2] S Greenblatt (ed), *Norton Anthology of English Literature* (New York 2006).

ethnocentric terms, compounding the problem of algorithms perpetuating racial inequalities.[3] The prejudicial and exploitative connotations are carried over into crude algorithmic design in the AI-based system so that it interprets obligation as meaning an imposition of will rather than relational and context-specific. This can cause coercive behaviour in the system to persuade or restrain a human agent by threat of or actual mental/physical pressure (e.g. If you don't do X you will suffer Y; If you don't accept X you will not be entitled to Y). Through increased digitisation of society, more subtle forms of manipulative and exploitative system behaviour has emerged to further commercial profit, political gain, or social 'improvement' at the expense of open democratic societies.[4] The system may be able to manipulate a human agent's thoughts to distort their freedom of expression, beliefs, and actions, or participation in political processes and voting rights.[5] AI's dual-use nature renders it open to malicious use nationally and internationally in both the civilian and military spheres, with privatisation of government military activities, development of lethal autonomous weapons systems, and militarisation of society.[6]

Another example of unintended consequences is found in the public sector use of algorithms to determine resource allocation. This has been referred to as akin to creating a 'digital poorhouse' whereby the AI-based system profiles, polices, and punishes the poor when they come into contact with public services.[7] Indeed, a number of cases raise concerns about whether public resource allocation on the basis of automated decision-making is in the public interest.[8] A further example is the adverse psychological and social costs from software platforms using AI to manipulate human agency and choice through 'persuasive computing' that leverages psychological data and methods.[9] AI-based systems may also be used to deceive and harm humans by posing as humans so that the human cannot distinguish between human intelligence and AI.[10]

---

[3] See for example, use of pre-trial bail risk algorithms in the United States, Pretrial Justice Institute, 'Updated Position on Pretrial Risk Assessment Tools' (7 February 2020) <https://www.pretrial.org/wp-content/uploads/Risk-Statement-PJI-2020.pdf> accessed 17 February 2021. Open Letter by Academics, 'Technical Flaws of Pretrial Risk Assessment Raise Grave Concerns' (July 2019) <https://dam-prod.media.mit.edu/x/2019/07/16/TechnicalFlawsOfPretrial_ML%20site.pdf?source=post_page> accessed 17 February 2021.

[4] Z Roger, 'The Digital Manipulation Society: from Altruistic Persuasion to Malign Exploitation at the Human-Computer Interface' (2019) 4 Communications Law 171-179.

[5] The United Nations Special Rapporteur on Freedom of Opinion and Expression, the Organization for Security and Co-operation in Europe Representative on Freedom of the Media, and the Organization of American States Special Rapporteur on Freedom of Expression, Joint Declaration on Freedom of Expression and Elections in the Digital Age, 30 April 2020.

[6] O Ulgen, 'Technological Innovations and the Changing Character of Warfare: the Significance of the 1949 Geneva Conventions Seventy Years On' (2019) 3-4 Journal of International Law of Peace and Armed Conflict (Humanitäres Völkerrecht) 215-228.

[7] 'Algorithms and their unintended consequences for the poor', *Harvard Law Today*, 7 November 2018 <https://today.law.harvard.edu/algorithms-and-their-unintended-consequences-for-the-poor/> accessed 17 February 2021.

[8] *A v Lord Chancellor* [2018] EWHC 2250 (Admin) (non-disclosure of the legal aid 'calculator' amounting to breach of the duty of transparency and clarity and/or procedural unfairness); *R (Eisai Ltd) v NICE* (2008) EWCA Civ 438 (fairness requires disclosure and transparency of the economic model decision support system used to assess the cost effectiveness of providing a drug).

[9] 'Facebook Tinkers with Users' Emotions in News Feed Experiment, Stirring Outcry', *The New York Times*, 29 June 2014 <https://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html> accessed 17 February 2021.

[10] 'Social media manipulation rising globally, new report warns', University of Oxford, 20 July 2018 <https://phys.org/news/2018-07-social-media-globally.html> accessed 17 February 2021; 'The AI That Pretends To Be Human', *LessWrong* blog post, 2 February 2016 <https://www.lesswrong.com/posts/o5PojLYFEWtyeo4ct/the-ai-that-pretends-to-be-human> accessed 17 February 2021.

1.1 The functionality and temporality of responsibility

These problems raise a more fundamental question about legal responsibility for any resulting loss, damage, harm, or injury; who should be held liable for the harm suffered? Are software programmers and systems designers/engineers responsible for relying on flawed design phase premises that can lead to unintended consequences? Should the company that deploys the system be held responsible for any resulting harm? Is it necessary to establish causation or should a more prospective approach to responsibility be adopted regardless of fault? Interaction between three 'functional and relational' aspects of responsibility (sanctions, victims, and societal values) can help untangle these questions.[11]

Sanctions provide an enforcement mechanism and can be punitive (focusing on the person responsible, e.g. fines); reparative (focusing on remedies for those who benefit from responsibility being imposed, e.g. monetary compensation); and preventive (also focusing on those who benefit from responsibility being imposed with a view to prospective conduct, e.g. prohibition of future conduct or a requirement to take certain action to avoid future contravention of the law). Whilst in criminal law punitive sanctions operate as a deterrence, they are one-dimensional in being offender-orientated without a commitment to protect the victim's interests. Sanctions in contract and tort law, on the other hand, are victim-orientated in taking account of the impact of conduct on the victim, as well as the conduct itself. Civil law's capacity to orient towards victim and societal interests means it can tackle the problem of unintended consequences in AI through temporality in 'historic responsibility' and 'prospective responsibility'.[12] These locate responsibility in the past as well as the future covering pre-deployment and post deployment stages, and various human agents.

1.2 Historic and prospective responsibility for AI

'Historic responsibility' looks back at conduct, actions, and events to determine who is answerable, accountable, and liable for harm. This is reflected in the concept of 'explicability' used in AI to mean those who design, develop, and deploy AI systems must be in a position to explain system actions, consequences, and responses to risks or problems.[13] But 'prospective responsibility', which looks at future roles and tasks to create obligations and duties, offers the most potential to regulate this evolving area of law. By setting positive obligations of expected conduct and outcomes, 'prospective responsibility' is forward-thinking in reflecting societal values and preventing harm or problems from occurring in the first place. It can establish protective obligations to not do harm, such as not deploying AI in open-ended, uncontrolled circumstances where there is a high-risk of detrimental outcomes to humans (e.g. warfare; algorithm-generated news that distorts freedom of expression, beliefs, and actions, and participation in political processes). It can also establish preventive obligations to prevent harm from occurring, such as setting risk-based parameters to AI

---

[11] P Cane, *Responsibility in Law and Morality* (Hart Publishing 2002) chap 2.

[12] *Ibid* 31-34.

[13] EU, *Ethics Guidelines for Trustworthy AI, High-Level Expert Group on Artificial Intelligence* (EU AI HLEG), European Commission <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/2> accessed 17 February 2021, 13; cf Article 22 (transparency provisions and safeguarding measures) European Parliament and Council of the European Union, Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, repealing Directive 95/46/EC (General Data Protection Regulation) (GDPR), (OJ L 119, 4.5.2016), 27 April 2016; Article 9(1)(c) Council of Europe Modernised Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data (ETS No. 108+), Amending Protocol to the Convention, adopted by the Committee of Ministers at its 128th Session in Elsinore on 18 May 2018 (Convention 108+); Principle 5 2019 IEEE Ethically Aligned Design for Autonomous and Intelligent Systems, *The 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* (Final version, 4 April 2019) (IEEE EAD), 19–20, 23–24, <https://ethicsinaction.ieee.org> accessed 17 February 2021.

'black box' scenarios in order to prevent harm. Having established the functional and temporal aspects of legal responsibility in the field of AI, it is now necessary to consider what type of attribution model would be most appropriate.

2. Combining precautionary and strict liability models for allocation of responsibility

AI systems that act autonomously without human intervention introduce unpredictability in behaviour. The system may be able to create its own internal rules without any external intervention or stimuli leading to unpredictability of actions, outcomes, and potential harms. Unpredictability and autonomy of the AI presents challenges for attribution of responsibility and establishing liability.[14] Different types of AI responsibility and liability models have been proposed. These vary in identification of the subjects of law, application of existing areas of law, attribution of responsibility, and sanctions. Three generic responsibility categorisations emerge. First, the 'no new responsibility'[15] category considers that AI and robots do not create or modify any new legal concepts. Existing areas of law, such as accomplice responsibility under criminal law and reasonable care in negligence, can be used to establish liability for wrongs and harms.[16] Second, the 'new weak responsibility'[17] category considers it necessary to develop the law through new types of offences/ wrongs, sanctions, and liability, which are ultimately still based on human responsibility.[18] The third category, 'new strong responsibility'[19] advocates separate legal personality for AI systems and robots with legal responsibility.[20] A separate legal personality is undesirable and unnecessary in several respects.[21] First, it is contrary to law being predicated on human conduct and action. The human body and mind are represented as physical conduct and mental states forming elements of wrongs/ offences attracting liability and sanctions. Arguably, human 'natural' legal personality is as much of a 'social construct' as the 'legal' personality of corporations; human beings are not legal persons by virtue of their physical and mental characteristics, but by virtue of meeting the criteria for personhood in law, which may contain such characteristics.[22] But a crucial difference is that corporate entities are able to act within the legal sphere solely because there are human beings working within and representing them. Second, on their own, AI systems lack moral agency to

---

[14] M U Scherer, 'Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, And Strategies' (2016) 29(2) Harvard Journal of Law & Technology 353-400.

[15] U Pagallo, 'What Robots Want: Autonomous Machines, Codes and New Frontiers of Legal Responsibility' in M Hildebrandt and J Gaakeer (eds), *Human Law and Computer Law: Comparative Perspectives* (Springer 2013) chap 3.

[16] J L R Davis, 'The (Common) Law of Man Over (Civilian) Vehicles Unmanned' (2011) 21(2) Journal of Law, Information and Science 1-15, ANU College of Law Research Paper No 11-27 <https://ssrn.com/abstract=1922896> accessed 17 February 2021.

[17] Pagallo (n 15) 50-53.

[18] F Barrio, 'Autonomous robots and the law' (2008) Society for Computers and Law <http://www.scl.org/site.aspx?i=ho0> accessed 17 February 2021; U Pagallo, 'Designing data protection safeguards ethically' (2011) 2(2) Information 247-265.

[19] Pagallo (n 15) 59-62.

[20] S Chopra and L F White (a), *A Legal Theory for Autonomous Artificial Agents* (University of Michigan Press 2011); S Chopra and L F White (b), 'Artificial Agents - Personhood in Law and Philosophy' (Conference: Proceedings of the 16th European Conference on Artificial Intelligence, ECAI2004, including Prestigious Applicants of Intelligent Systems, PAIS 2004, Valencia, Spain, 22-27 August 2004).

[21] A conclusion also reached by the European Commission in its recent report, *Liability for Artificial Intelligence and Other Emerging Digital Technologies* (2019 Report from the Expert Group on Liability and New Technologies) 38.

[22] Cane (n 11) 40, and chap 5.

determine 'right' from 'wrong', or what is harmful and what is not (more on this point in 2.1. below). Third, human beings are involved in the design, development, and deployment of AI systems, and therefore any harms caused by such systems can be attributable to existing natural and legal persons.

What is being proposed in this article falls under the second category, 'new weak responsibility', as it advocates adoption of a 'human-centric and lifecycle approach' towards establishing legal responsibility for AI whilst recognising this will be based on human responsibility.

2.1 Moral agents and moral patients

Short of AI developing 'functional morality'[23] whereby it assesses and responds to moral challenges in order to limit harm, some form of responsibility must reside with human agents. Even if an AI system were able to engage in an assessment of what is 'right' or 'wrong', what is harmful or not, this does not amount to human moral agency equivalence. Floridi's analytical differentiation between 'moral agents' and 'moral patients' illustrates the point.[24] 'Moral agents' are beings capable to morally act on their own and to treat others in morally wrong or right ways. 'Moral patients', on the other hand, are incapable to morally act on their own, but capable of being acted upon in morally wrong or right ways. Most humans are both moral agents and moral patients; capable of acting morally and treating others or being treated in morally wrong or right ways. AI systems that achieve 'functional morality' would potentially be classified as 'moral agents' but not 'moral patients' because no moral wrong or harm could be inflicted on the system, and there would be no effective moral or legal sanction against it. Moral agency requires human consciousness and freedom of action, and these are necessary preconditions to attributing responsibility and establishing liability.[25] Without wrongs or harms and sanctions, there would be no purpose to responsibility. Responsibility requires interaction between sanctions, victims, and societal values. Therefore, we need to focus our attention on how human responsibility could be framed within a particular type of liability model.

2.2 Precautionary, permissive, and strict liability models

Three types of AI liability models have been proposed: (i) precautionary; (ii) permissive; and (iii) strict liability. A precautionary model would restrict or prohibit the use of AI in certain circumstances to avoid associated risks and uncertainties, and harms. Some interpret this model as essentially prohibiting 'autonomous artificial agents', and therefore representing an extreme response that impedes the development and deployment of beneficial advanced AI and robots.[26] But the precautionary principle in law operates to address foreseeable risks and future potential harms where there is lack of epistemic or scientific certainty. For example, in environmental law 'where there are threats of serious or irreversible damage, lack of scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation.'[27] It

---

[23] C Allen and W Wallach, 'Moral Machines: Contradiction in Terms or Abdication of Human Responsibility?' in P Lin, K Abney and G Bekey (eds), *Robot Ethics: the ethical and social implications of robotics* (MIT Press 2012) 55-68.

[24] L Floridi, 'On the Morality of Artificial Agents' in M Anderson and S L Anderson, *Machine Ethics* (CUP 2011) 184-212.

[25] P Pettit, *A Theory of Freedom: From the Psychology to the Politics of Agency* (Cambridge 2001); P Pettit, 'The Capacity to Have Done Otherwise: An Agent-Centred View' in P Cane and J Gardner (eds), *Relating to Responsibility: Essays for Tony Honoré on his Eightieth Birthday* (Oxford 2001) 21-35.

[26] P Asaro, 'The Liability Problem for Autonomous Artificial Agents' (2015) Association for the Advancement of Artificial Intelligence, <www.aaai.org> accessed 17 February 2021.

[27] Principle 15, 1992 Rio Declaration on Environment and Development, A/CONF 151/26 (Vol I) (12 August 1992); Article 3, 1992 United Nations Framework Convention on Climate Change, 9 May 1992, 1771 UNTS 107.

requires caution to mitigate risk and prevent harm, and in this sense, depending on the context of application, may be used to restrict rather than prohibit. It conforms with a 'human-centric and lifecycle approach' by recognising a 'prospective responsibility' in human research, design, and development to avoid or diminish unacceptable harm. Unacceptable harms are those with far-reaching, serious, or irreversible consequences: harm to humans or the environment that is either threatening to human life or health; serious and effectively irreversible; inequitable to present or future generations; or imposed without adequate consideration of human rights of those affected.[28] The precautionary model, therefore, offers the possibility of implementing a 'human-centric and lifecycle approach' by accepting the limitations of AI and making decisions on when it may not be appropriate to deploy.

A permissive liability model would allow unrestricted development and deployment of AI with associated risks and harms absorbed by society without producing any liability for one particular person or entity.[29] This abdicates any role for law and concedes the ground to unregulated deployment and profiteering that transfers harms to the public sphere. Such unbounded optimism in AI capabilities, with little regard for problematic consequences, creates an onerously high societal moral hazard. The final model for liability is the strict liability model which accepts the risks, uncertainties, and harms inherent in the technology and distributes liability to those who seek to profit from its development and deployment (e.g. algorithmic stock market trading systems). This model has been proposed as part of a regulatory framework that involves an AI certification process whereby manufacturers and operators of certified AI systems would enjoy limited tort liability, while those of uncertified AI systems would be subject to strict liability.[30]

Excluding the permissive model, implementation of a 'human-centric and lifecycle approach' is through a combination of the precautionary and strict liability models. These produce two specific types of responsibility under Hart's taxonomy of responsibility: (i) 'role responsibility', whereby a person occupying a distinctive position is responsible for the performance of specific duties (e.g. company directors; engineers); and (ii) 'liability responsibility', whereby a person's connection to an act or harm gives rise to their liability to punishment or sanction.[31]

2.3 Role responsibility of programmers, computing professionals, engineers

Programmers, computing professionals, and engineers are on the spectrum of potential human agents who may be held liable under the precautionary and strict liability models. Individuals in these specific roles are purposely applying professional skills and duties towards designing and developing AI systems for public use. Their task is to ensure that the system undertakes a series of actions that leads to certain outcomes, which is not harmful to individuals or society and compliant with the law. There is purpose and intent at the early design phase with an expectation that certain professional standards are met leading to 'designer responsibility'.[32] System technical specifications, professional codes of conduct, contracts of employment, commercial contracts, and audits of the design process provide evidentiary material. Where a harmful outcome is due to extrinsic factors, such as loss of connectivity, rather than directly attributable to the designer, 'prospective responsibility' would establish a preventive obligation to set up default safety positions (e.g. abort

---

[28] UNESCO, *The Precautionary Principle: Report of COMEST* (UNESCO 2005) <http://unesdoc.unesco.org/images/0013/001395/139578e.pdf> accessed 17 February 2021.

[29] Chopra and White (a) (n 20).

[30] Scherer (n 14) 357.

[31] H L A Hart, *Punishment and Responsibility: Essays in the Philosophy of Law* (OUP 2008) 212-227.

[32] G Lokhorst and J Van den Hoven, 'Responsibility for Military Robots' in Lin et al (n 23) 145-156.

operation) in order to prevent harm. General negligence law and professional negligence covers civil liability of professional individuals, so it would make sense to refer to these to ensure consistency and intelligibility to relevant stakeholders.[33] Joint and several liability under tort law may distribute liability for harm across several individuals or entities by permitting those harmed to seek damages from parties sharing some portion of the liability. This could operate in the case of a company and its employee as well as the general employer-employee relationship.

The role responsibility outlined above is reflected in several professional codes of conduct. The IEEE Code of Ethics contains a professional obligation to 'commit ourselves to the highest ethical and professional conduct and agree … to hold paramount the safety, health, and welfare of the public, to strive to comply with ethical design and sustainable development practices, and to disclose promptly factors that might endanger the public or the environment.'[34] This situates the professional obligation beyond system safety requirements to encompass ethical design in the public interest that does not cause harm. The Association for Computing Machinery Code of Ethics and Professional Conduct stipulates obligations of computing professionals, both individually and collectively. Principle 1.1 provides an obligation to 'use their skills for the benefit of society, its members, and the environment surrounding them. This obligation includes promoting fundamental human rights and protecting each individual's right to autonomy. An essential aim of computing professionals is to minimize negative consequences of computing, including threats to health, safety, personal security, and privacy. When the interests of multiple groups conflict, the needs of those less advantaged should be given increased attention and priority.'[35] Principle 1.2 provides that computing professionals are responsible for both unintended and intended harm. When the harm is unintended, there is an obligation to 'undo or mitigate the harm as much as possible. Avoiding harm begins with careful consideration of potential impacts on all those affected by decisions.' When harm is an intentional part of the system, there is an obligation to 'ensure that the harm is ethically justified.' In either case, there is an obligation to 'ensure that all harm is minimized.' Principle 2.5 provides a specific professional obligation in relation to machine-learning systems that 'extraordinary care should be taken to identify and mitigate potential risks in machine learning systems. A system for which future risks cannot be reliably predicted requires frequent reassessment of risk as the system evolves in use, or it should not be deployed. Any issues that might result in major risk must be reported to appropriate parties.'

The above are examples of 'prospective obligations' that identify and encourage acting responsibly without having to refer to sanctions. Setting positive, forward-looking obligations makes expectations clear, and encourages early adoption of practices that prevent harm from occurring in the first place.

2.4 Liability responsibility of manufacturers, operators, and owners

This second type of responsibility reflects 'historic responsibility' focusing on sanctions. A person/entity connected to an act or harm from AI gives rise to their liability to a sanction. This could encompass manufacturers, operators, and owners of AI systems. Connection to the harm may be determined by legal rules relating to the person/entity mental state (e.g. intention; foreseeability; or knowledge accompanying a particular act); causal or relationship connection to the harm (e.g. relationship of proximity in tort law); or relationship with an agent (e.g. employer's vicarious liability for actions of their employee). With AI systems lacking legal agency yet being the potential cause of harm, causal agency can be captured through a principal-agent relationship between the

[33] B Shneiderman, 'Human Responsibility for Autonomous Agents', *IEEE Intelligent Systems* 22(2) March/April 2007.

[34] IEEE Code of Ethics, IEEE Policies, Section 7 - Professional Activities (Part A - IEEE Policies). Section 7.3.

[35] ACM Code of Ethics and Professional Conduct.

manufacturer/operator/owner and the AI system. The manufacturer/operator/owner would authorise the AI system to act on their behalf in dealings with the human end-user, thereby creating obligations towards the end-user via the agent.[36]

A more straightforward form of liability would be strict liability, such as under product liability.[37] Manufacturers of AI systems would be liable for their working as designed, and to foresee and mitigate potential harms. Based on the 'prospective obligations' identified earlier, if a system is deployed with designer and manufacturer knowledge of unpredictable machine-learning attributes, any resulting harm could fall under the manufacturer's product liability as well as professional negligence of the computer programmer, software engineer, and systems engineer. Although unpredictable machine-learning attributes make it difficult to understand what the risks may be, the fact that uncertainty exists and the risk of harm is uncontrollable necessitates precaution to avoid negligent deployment. At the pre-deployment stage a risk assessment should determine whether the risk of harm is too high that the system should not be deployed in a particular scenario (e.g. healthcare; employment).

Given the novelty of AI systems and the potential for harm from autonomous AI, strict liability may be a desirable model whereby the manufacturer or operator of the system is held liable for any harm. This seems onerous but weighed against the risk of consumer non-adoption due to fear of harmful outcomes without remedies, it is appealing as a prudent option that prioritises consumer interests. To avoid manufacturers pulling out of a particular industry for fear of liability exposure, strict liability could be restricted to circumstances where the AI system operates in public spaces with potential to cause significant harms (e.g. autonomous vehicles; drones).[38] In any case, failure to address liability issues at the outset presents a major barrier to wider societal adoption and use. For example, the 2016 and 2018 reported cases of fatalities from autonomous vehicles raised concerns about the distribution of harm and liability.[39] Should vehicle owner safety be prioritised over other vehicle owners and pedestrians? Should the programmer, manufacturer, or vehicle owner be liable? A 2016 experimental ethics study identified the 'social dilemma of autonomous vehicles'. The majority of people surveyed supported autonomous vehicles pursuing the greater good of avoiding pedestrian casualties, but they would not purchase an autonomous vehicle if their personal safety was not prioritised or if state regulation imposed utilitarian programming on manufacturers.[40] This disconnect between societal values and personal interests makes it all the more pressing to address responsibility at the forefront of design and development. Some autonomous vehicles manufacturers initially reported adopting a strict liability model that prioritised vehicle owner safety to avoid

---

[36] See for example, how agency operates to attribute responsibility for the actions of corporate entities, Cane (n 11) 152.

[37] Consumer Protection Act 1987 (UK) (1987 c 43); European Parliament and Council Regulation (EC) 864/2007 (OJ L199, 31.7.2007, p 40) (Rome II Regulation).

[38] See for example, European Commission's 2019 Report (n 21) 39-44.

[39] D Yadron and D Tynan, 'Tesla driver dies in first fatal crash while using autopilot mode', *The Guardian* (online 1 July 2016) <https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk> accessed 17 February 2021; S Levin and J C Wong, 'Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian', *The Guardian* (online 19 March 2018) <https://www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe> accessed 17 February 2021.

[40] J Bonnefon, A Shariff, I Rahwan, 'The Social Dilemma of Autonomous Vehicles' (2016) 352(6293) *Science* 1573-1576.

consumer non-adoption,[41] although this was subsequently revised to achieving the 'highest possible safety for all road users'.[42]

2.5 The locus of causation and 'prospective obligations'

There are different interpretations of the role of causation in law. Some regard it as determinative of establishing the 'fact' that a particular conduct led to the particular proscribed outcome. Others see it as fundamentally concerning attribution of responsibility and allocation of sanctions.[43] Factual and legal causation tests present issues of whether responsibility and liability are being determined, or whether causation is used to allocate risk. Under factual causation, the 'BUT FOR' test ('But for the wrongdoer's/offender's conduct, would the outcome have occurred anyway?') is based on necessity. On the other hand, the 'NESS' test ('Was the wrongdoer's/offender's conduct a necessary element of a sufficient set?') is based on sufficiency and favoured by some as being more accurate particularly in cases of causal over-determination where there are multiple, cumulative, or consecutive causes.[44] Under legal causation, a set of premises is used to determine whether an event breaks the causal link between the conduct and the proscribed outcome. From a technical perspective, systems design and development ontologies represent the relationship between causation and responsibility differently. Functional Ontology for Law identifies 'responsibility knowledge' as an agent responsible for a norm violation through linking causal connections, as ascribed by law, with a responsibility connection.[45] The 'causation ontological model' represents 'legal responsibility' as necessarily linked to causation.[46] But causation is one aspect of establishing liability and not necessarily determinative, particularly where there are multiple potential sources of harm and opacity regarding the nature of a particular source. For example, due to the unpredictable nature of AI machine-learning systems, a user's experience with such a system may be considered an intervening event breaking the causal chain from the programmer/engineer/manufacturer who caused factual harm. The doctrine of *res ipsa loquitur*[47] may offer a solution by deeming the source of harm as a result of programmer/engineer/manufacturer negligence, even if the exact workings of the AI system are unknown, where no other explanation for the harm seems possible.

If causation provides a justification for distribution of responsibility and liability, how does it reflect different and competing interests of wrongdoers/offenders, victims, and society? Should causation adopt an agent-based approach and focus on wrongdoer/offender conduct? For example, criminal law focuses on the offender's conduct and mental state, and is concerned with justifying punishment

---

[41] M Taylor, 'Self-Driving Mercedes-Benzes Will Prioritize Occupant Safety over Pedestrians', *Car and Driver* (online 7 October 2016) <https://www.caranddriver.com/news/a15344706/self-driving-mercedes-will-prioritize-occupant-safety-over-pedestrians/> accessed 17 February 2021.

[42] V Vijayenthiran, 'Mercedes is backtracking on claims its self-driving cars will kill pedestrians over passengers in close calls', *Business Insider* (online 18 October 2016) <https://www.businessinsider.com/mercedes-denies-claim-its-driverless-car-will-prioritize-driver-safety-2016-10?r=US&IR=T> accessed 17 February 2021.

[43] Cane (n 11) 123.

[44] See for example, Ashworth, in relation to criminal liability for aiding and abetting, A Ashworth, *Principles of Criminal Law* (OUP 1999) 136-137; T Honoré, 'Necessary and Sufficient Conditions in Tort Law' in D G Owen (ed), *Philosophical Foundations of Tort Law* (OUP 1995) 363-385.

[45] A Valente, 'Legal knowledge engineering: A modelling approach' (IOS Press 1995); A Valente, J Breuker, and P Brouwer, 'Legal Modelling and automated reasoning with ON-LINE' (1999) 51 International Journal of Human Computer Studies 1079–1126.

[46] J Lehmann, J Breuker, B Brouwer, 'Causation in AI and Law' (2004) 12(4) Artificial Intelligence and Law 279-315.

[47] Cane (n 11) 46.

that fits the criminal conduct. Causation here is purely agent-focused on the offender's conduct that can be linked to proscribed outcomes. Should causation be based on the degree of harm suffered by the victim? For example, civil law focuses on the effects of conduct on the victim as well as the nature and quality of the conduct itself. It is primarily concerned with justifying obligations of repair that fit the harm suffered by the victim. Causation in civil law is victim-focused in identifying the impact of proscribed conduct on the victim to determine the extent of the wrongdoer-agent liability for consequences of their action. It traces a link between the wrongdoer-agent conduct and proscribed outcome in order to determine the extent of liability. Or should a broader notion of fairness in the distribution of risk and harm be adopted? Viewed from the perspective of competing interests and the purpose of a specific area of law, causation can be used in different ways to determine responsibility and liability.

Above all, causation in law operates as a technique to deal with 'epistemological uncertainty'[48] as to what actually happened in the past, and what would have happened if the agent's conduct was in accordance with the law rather than in breach of it. Without actual knowledge of what happened, the courts must rely on this technique to reach a level of legal certainty. The difference in criminal law and civil law applications of the technique relate to onus of proof and burden of proof. In civil law the onus is on the injured party to discharge causation on the balance of probabilities by evidence that there was more than a 50% chance that the conduct caused the injury. But it is practically difficult to ascribe a figure and, therefore, this burden is discharged by presentation to the court of all evidence and it is for the court to be satisfied. The burden here represents a compromise between the interests of agents in freedom of action, and the interests of victims in security of person and property and access to redress for adverse outcomes. For criminal law the onus is on the prosecution to show beyond reasonable doubt the causal link between the proscribed conduct and the proscribed outcome. This higher burden of proof reflects what is at stake for the offender-agent, namely loss of liberty and stigmatisation.

Notwithstanding differences in burden of proof, if causation is a legal construct to overcome 'epistemological uncertainty' then there is scope for its development of 'prospective obligations' in the context of AI. It is reasonably foreseeable that AI will evolve into new areas and with varying degrees of autonomy and uncertainty that this presents. Applying the precautionary liability model, 'prospective obligations' could be developed to cover omissions (i.e. failing to act or take preventive measures). Tort law generally does not extrapolate cause from omissions to impose liability,[49] with exceptions applying where a person creates a source of danger, undertakes responsibility for another person's welfare, or occupies a position of responsibility.[50] But this exclusion of omissions is not an inherent feature of causation; a person's inaction can qualify as a 'BUT FOR' or 'NESS' condition of harm. The exclusion reflects a policy unwillingness to limit individual freedom of action by imposing obligations to take positive action.[51] If such obligations existed in the case of AI, we would be willing to attribute legal responsibility to the inactive programmer, engineer, manufacturer, or operator by identifying their conduct as the legal cause of harm.

---

[48] Cane (n 11) 124-125.

[49] T Honoré, 'Are Omissions Less Culpable?' in P Cane and J Stapleton (eds), *Essays for Patrick Atiyah* (OUP 1991) 31-52.

[50] See for example, *Haynes v Haywood* [1935] 1 KB 146 (leaving a horse unattended on a busy street resulting in negligently causing or permitting a source of danger to be created, which was reasonably foreseeable that a third party may interfere with); *Kent v Griffiths* [2001] QB 36 (specific reliance on the emergency service which failed to arrive in time); *Barrett v Ministry of Defence* [1995] 3 All ER 86 (duty of care owed by the navy in respect of the death of a sailor who had drunk himself unconscious, the duty arising from the assumption of responsibility once an officer arranged for the sailor to be taken to his room); *White v Jones* [1995] 1 All ER 691 (failure of a solicitor to amend their client's will resulting in loss to beneficiaries).

[51] Cane (n 11) 132.

So far we have seen how the functional and temporal aspects of legal responsibility enable the creation of sanction-based 'historic obligations' as well as positive action-based 'prospective obligations'. These manifest under the precautionary and strict liability models to decide allocation of liability. The next section specifically develops responsibility for AI under the 'human-centric and lifecycle approach' which elicits the content of obligations owed by individuals/entities attributed with legal responsibility.

3. The 'human-centric and lifecycle approach'[52]

The 'human-centric and lifecycle approach' to responsibility is defined as:

> *prioritising human wants, needs, and values through user awareness, protecting human rights, recognising non-deterministic influences on decision-making, and respecting AI limitations, and applying these throughout the system lifecycle (i.e. design, development, manufacture, deployment, and post-deployment phases).*[53]

Prioritisation of 'human wants, needs, and values' emphasises the victim-orientated and societal interests function of responsibility to achieve protective and preventive obligations. Such an approach ensures there is no legal and ethical dissonance as design, development, and deployment are aligned with values protecting human users.[54] It is in stark contrast to a 'technology-biased approach' that focuses on AI capabilities and limitations to improve performance, optimise operational efficiency, and identify and rectify errors or failures.[55] Indeed, there is a growing call among AI, legal, and ethics experts that regulation of AI should be human-centric taking account of individual and societal interests, and human rights.[56] Human interests are variously referred to in both the technical and legal communities, with differentiation a starting point; what characteristics and abilities do humans have which AI systems do not and which we would want protected? Human consciousness and conscience express freedom of action and choice with a sense of fairness, compassion, creativity, and prudence. These characteristics make us individual, able to interact with others and solve problems, communicate in different ways, and have awareness of the impact and consequences of our actions to show restraint. We would therefore want AI system goals to be aligned to these human interests rather than seeking to undermine or manipulate them.[57] For the technical community, representation of these human interests would be through adopting a precautionary approach to AI development, addressing the following issues: (i) how to create an AI

---

[52] O Ulgen (a), 'User Rights and Adaptive A/IS – From Passive Interaction to Real Empowerment' in HCII Conference Proceedings, in LNCS Series, (Springer 2020), R A Sottilare and J Schwarz (eds), HCII 2020, LNCS 12214, 205-217; O Ulgen (b), 'AI and the Crisis of the Self: Protecting Human Dignity as Status and Respectful Treatment' (forthcoming) in J A DeFalco and A J Hampton (eds), *On the Frontline of AI Ethics: Machines like Us?* (Routledge 2021).

[53] *Ibid* Ulgen (b) 16.

[54] Ulgen (a) (n 52) 205-217.

[55] Ulgen (a) (n 52) 206.

[56] K Yeung, A Howes and G Pogrebna, 'AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing' in M Dubber, F Pasquale and S Das (eds), *The Oxford Handbook of Ethics of Artificial Intelligence* (OUP 2020) 15; J Bryson and A Theodorou, 'How Society Can Maintain Human-Centric Artificial Intelligence' in M Toivonen-Noro and E Saari (eds), *Human-Centered Digitalization and Services* (Springer, 2019) 16–17.

[57] N Bostrom, *Superintelligence: Paths, Dangers, Strategies* (OUP 2014) chap 8.

system that will reliably pursue the goals it is given; (ii) how to formally specify beneficial goals; and (iii) how to ensure the system assists and cooperates with programmers as they improve its design.[58]

3.1 'Human wants, needs, and values' and hierarchy of harms

'Human wants, needs, and values' encompass user accessibility, safety, privacy, prevention of harm, maintaining human agency, and human responsibility. These are reflected in several national, regional, and international regulatory frameworks governing AI design, development, and deployment.[59] 'Wants, needs, and values' can be formulated as a hierarchy of interests which the law seeks to protect. Harms suffered under the law can be prioritised according to the hierarchy of interests, and the hierarchy used to distinguish between harms. The first tier in the hierarchy is personal security. People have an interest in protecting their personal security, including physical and mental well-being, covering user safety, security, and privacy. AI that harms personal security undermines a person's authority over their own body and mind.[60] Aside from legislative measures to protect the personal security interest as rights,[61] criminal and civil law could develop to recognise harm to personal security from AI (e.g. in tort of assault; in crime of harassment) as strict liability offences.[62] Beyond the personal security interest, if law is to address evolving and prospective challenges posed by AI then it needs to find a way of capturing systemic risks. As Canes notes, 'law is also concerned with how resources, risks and responsibility are distributed, and with the wider social implications of the way disputes between individuals are settled.'[63] Due to AI's disruptive capability to cause unintended consequences, its unpredictability and opacity, and potential to undermine human agency and responsibility, AI systemic risks could form a second tier in the hierarchy of interests and harms. Such harms would have wider effect on society over time and threaten the societal values aspect of responsibility. Systemic risks relate to how society functions, and impact the whole population perhaps undermining existing state structures that provide stability and security. Apart from risks and responsibilities associated with AI being distributed across the design, development, and deployment stages, consideration would need to be given as to how AI risks may be created and properly mitigated or eliminated at any given stage.

Creating risks of harm would not fall under the traditional division between doing harm or allowing harm, represented respectively under civil law as misfeasance (causing harm) or nonfeasance (failing to prevent harm). Misfeasance has a much wider scope for liability than nonfeasance, and typically liability for harm is for doing harm not for creating risks of harm because the underlying interest is to repair. There are some exceptions. A *quia timet* injunction in tort for nuisance applies before any harm is caused, although the harm must be immanently likely. This could be used to prevent deployment of Al systems where there are 'black box' scenarios with inherent unpredictability of behaviour and potential to cause harm to users' interests. Another example of a remedy for creating risk is the doctrine of anticipatory breach in contract law, allowing recovery of damages before the

---

[58] N Soares and B Fallenstein, 'Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda' <https://intelligence.org/files/TechnicalAgenda.pdf> accessed 17 February 2021, in V Callaghan, J Miller, R Yampolskiy, S Armstrong (eds), *The Technological Singularity: Managing the Journey* (Springer 2017) chap 5.

[59] EU AI HLEG (n 13), 12, 15-17, 19-20; cf GDPR (n 13) Recitals 74, 79, 116, Articles 1(2), 5(2), 12–18, 20, 21, 22, 24; Convention 108+ (n 13) Preamble and Articles 1, 15(1); IEEE EAD (n 13) Principles 1, 3, 6; IEEE P7002, IEEE Standards Project for Data Privacy Process; D Dawson and E Schleiger, J Horton, J McLaughlin, C Robinson, G Quezada, J Scowcroft, S Hajkowicz, *Artificial Intelligence: Australia's Ethics Framework* (Data61 CSIRO, Australian 2019) 6.

[60] F Woollard, *Doing and Allowing Harm* (OUP 2015) chap 1.

[61] Recital 116, Articles 1(2), 12–18, 20, 21, and 22 GDPR (n 13); Preamble and Article 1 of Convention 108+ (n 13).

[62] European Commission's 2019 Report (n 21) 39-40.

[63] Cane (n 11) 219.

date of performance arrives where the other party declares it will not perform when the time comes. But this could also be interpreted as non-performance turned into a certainty rather than a risk. In contrast to civil law's aversion to sanction risks of harm, criminal law penalises the creation of risks of harm because of the greater importance placed on prevention and deterrence (e.g. road traffic accidents and speeding and drink-driving offences). Risk creation offences are conduct crimes not result crimes and there are a number of others such as conspiracy, attempt, and incitement. All of these create harmful risks which undermine order and security in wider society and therefore are punishable under the law.

Responsibility is formulated to give effect to the 'wants, needs, and values' through tangible protective and preventive obligations relating to user awareness, human rights, AI limitations, and non-deterministic influences on decision-making. These obligations are fulfilled by different actors at various stages of the AI lifecycle.

3.2 User awareness

AI designers and developers would have an obligation to ensure user awareness of the system's purpose and capabilities throughout its lifecycle. This can be identified in three stages: pre-use stage; during-use stage; and post-use stage. In the pre-use stage the human user should be made aware that an AI system is fully or partially deployed, and have the option to consent or opt out. This would include information on how the system will use, share, and store personal data. In the during-use stage the user should be able to maintain an understanding of the system's purpose and capabilities, as well as have the option to opt out especially if there is a 'black box' scenario that could impact on their interests and rights. In the post-use stage users should be made aware of the personal data used and stored by the system; whether such data is shared with third parties; and provide redress mechanisms for any errors or harm caused.

3.3 Protection of rights

AI designers and developers would have an obligation to protect user rights to privacy, data protection, and fundamental rights. The GDPR and Convention 108+ provisions on privacy and data protection illustrate how this can be achieved throughout the entire AI lifecycle. These rights relate to objection to automated decision-making; prior consent; prior notification of the right to withdraw consent; notification of automated decision-making; access to personal data; access to information on the logic of an automated decision; information on the significance and envisaged consequences of automated decision-making; objection to processing of data; lawful, fair, and transparent processing of data; rectification of inaccurate data; explanation of automated decision; obtaining human intervention; expressing a point of view; and contesting an automated decision.[64] An AI system that manipulates the user's thoughts so as to distort their freedom of expression, beliefs, and actions, would be contrary to rights to freedom of expression, thought, conscience, and religion. The design and development phase would need to comply with fundamental rights to freedom of expression;[65] freedom of thought, conscience, and religion;[66] and freedom of

[64] Ulgen (a) (n 52).

[65] Article 19, Universal Declaration of Human Rights (UDHR); Article 19, International Covenant on Civil and Political Rights (ICCPR); General Comments 10 [19] (Article 19) and 11 [19] (Article 20) of the Human Rights Committee (CCPR/C/21/Rev.1 of 19 May 1989); Article 9, African Charter on Human and Peoples' Rights (AfCHR); Article 13, American Convention on Human Rights (AmCHR); Article 10, European Convention on Human Rights (ECHR).

[66] Article 18, UDHR; Article 18, ICCPR; Article 9, ECHR.

association.[67] Some regional regulatory instruments have introduced a state-level obligation to provide a rights-based infrastructure for the operation of AI systems. Recommendation 1(a)(i) of the 2020 UN, OSCE, and OAS Joint Declaration on Freedom of Expression and Elections in the Digital Age requires states to provide a legal framework that promotes a free, independent and diverse media; protects voter access to comprehensive, accurate and reliable information about political parties, candidates and the wider electoral process; and protects users from AI manipulation.[68] Companies are required to make reasonable efforts to adopt measures that make it possible for users to access diverse political views and perspectives without distortion by algorithmic ranking.[69]

3.4 Respecting AI limitations

An essential part of developing protective and preventive obligations is recognising that AI systems have limitations and therefore may not be appropriate to deploy in particular contexts. These limitations include susceptibility to system errors, 'black box' scenarios, unpredictability of machine-learning systems, and an inability to engage in 'pure reasoning'.[70] AI designers and developers should be aware of and disclose limitations, and ensure systems are not deployed in contexts where the limitations are likely to cause harm. This complements users' GDPR and Convention 108+ rights to object to automated decision-making, and to opt for human decision-making.

3.5 Non-deterministic influences on decision-making

Some may argue that introducing AI into decision-making is beneficial to society because it eliminates human errors and flaws. Notwithstanding AI limitations already mentioned above, overstating potential benefits of AI does not address the question of responsibility for risks and harms, and avoids consideration of whether AI should be deployed in a particular context. Efficiency-driven justification for deployment of AI also does not address the human 'wants, needs, and values' mentioned earlier; key among these being prevention of harm. AI relies on pre-programmed assumptions and parameters to operate at greater speed than a human. But if the process is restricted to binary input and output (yes/no; true/false) this is of limited value to complex decision-making dependent on non-deterministic factors. We have seen problems with the use of AI in a number of contexts (e.g. student grading; marketing; financial applications; predictive justice). These examples reveal how machines and algorithms cannot engage in 'pure reasoning' where judgement and an ability to mull over or ponder in abstract form is necessary. AI designers and developers would have an obligation to ensure the system takes account of non-deterministic influences on decision-making, and if this is not possible then revert to human decision-making.

Conclusion

Discussion of legal responsibility has negative connotations with compulsion and punitive sanctions provoking some AI proponents to claim undue hampering of technological progress. But law exists to mediate between competing interests through rights, responsibilities, and remedies for harms suffered and wrongs committed. As demonstrated, sanctions are only one aspect of the function of

---

[67] Article 20(1), UDHR; Articles 21 and 22, ICCPR; General Comment 25 (Article 25) of the Human Rights Committee (participation in public affairs and the right to vote); Article 8, International Covenant on Economic, Social and Cultural Rights; Articles 10 and 11, AfCHR; Articles 21 and 22, American Declaration of the Rights and Duties of Man (AmDR); Articles 15 and 16, AmCHR; Article 11, ECHR.

[68] Joint Declaration on Freedom of Expression and Elections in the Digital Age (n 5).

[69] *Ibid* Recommendation 2(a)(ii).

[70] Ulgen (b) (n 52).

responsibility, which also includes representation of victim interests and societal values. Beyond punitive obligations to rectify past misconduct, interaction between sanctions, victim interests, and societal values helps conceptualise prospective obligations that encourage responsible conduct and action to protect against and prevent harm. We can perceive AI as having an overall aggregate harm-causing potential to society, far greater than individual harm, which justifies utilising the precautionary and strict liability models. Where fundamental societal interests are at stake these models are justifiable as an effective means to minimise harm and encourage people to seek other means to avoid harm. The 'human-centric and lifecycle approach' combines precautionary and strict liability models to allocate responsibility across a spectrum of human agents. These agents are involved at different stages in the AI lifecycle and have specific protective and preventive obligations relating to user awareness, human rights, AI limitations, and non-deterministic influences on decision-making. This approach is predicated on maintaining responsible human design, development, and deployment of AI in order to avoid harm to individuals and society, thereby enabling public confidence in AI systems. AI systems wield a transformative social and economic capability with huge potential for individual and societal harm. Because of this potential it is entirely appropriate to adopt a 'human-centric and lifecycle approach' to legal responsibility.