The Law and Ethics of Artificial Intelligence: A Case Study into the Unique Challenges of Military Autonomous Machines

By

Jennifer Adams

STUDENT DECLARATION FORM

Concurrent registration for two or more academic awards
I declare that while registered as a candidate for the research degree, I have not been a registered candidate or enrolled student for another award of the University or other academic or professional institution.

Material submitted for another award
I declare that no material contained in the thesis has been used in any other submission for an academic award and is solely my own work Collaboration Where a candidate's research programme is part of a collaborative project, the thesis must indicate in addition clearly the candidate's individual contribution and the extent of the collaboration. Please state below: (N/A)

Signature of Candidate ____

Type of Award: Doctor of Education
School: school of Law and Policing

Abstract

Mankind has a long history in the innovation and use of tools to make work and life easier, plus dominate in times of war. Artificial intelligence (AI) started out in the 1950's as just another tool in mankind's toolbox, however, the trajectory to date has advanced AI to a point where independent thought by autonomous machines (AMs) is a reality and where this thesis argues machine consciousness will emerge within our lifetime. The thesis explores the emerging intersection of machine consciousness, ethics, and International Humanitarian Law (IHL) in the context of military autonomous machines (MAMs). The research critically examines whether current legal and ethical frameworks can accommodate the development of conscious MAMs, addressing the critical issues of accountability, liability, and ethical obligations.

The thesis identifies gaps in legal provisions and proposes the necessity of recognising MAMs as legal persons, thereby assigning rights and duties to mitigate liability concerns. In doing so, the thesis identifies the True Value Alignment Problem (TVAP), asserting that beyond aligning MAMs with our values under IHL, the challenge lies in extending IHL principles to MAMs, thus ensuring ethical treatment and protection for their 'life'. Furthermore, it is remined throughout this thesis that Military AMs (MAMs) will not choose their path in life and will be 'switched on' in the battlefield, therefore we must be evermore rigorous when deploying MAMs. The thesis also calls for the development of robust regulatory frameworks to govern the deployment, liability, and ethical considerations of MAMs, ensuring compliance with IHL. Recognising MAM autonomy and their potential moral agency is essential to maintaining legal clarity and parity, alongside preventing ethical and accountability gaps in future warfare.

This study contributes to the broader discourse on AI and IHL, by emphasising the crucial need for legal adaptation and ethical foresight, in the development and deployment of conscious MAMs.

Table of Contents

List of Figures

Acknowledgements

Thank you to everyone who has supported me through my thesis; supervisors, experts, colleagues and friends. I am especially grateful to my husband, who selflessly paused his life so I could fulfil my ambition, and to my children and grandson who have only ever known me on this journey.

I thank you all

1. Introduction

Our stable legal and ethical landscape, where we know our rights and duties, is argued here is about to be disrupted by the development of conscious autonomous machines (AMs) and military autonomous machines (MAMs). For the purposes of this thesis, AMs, including MAMs, refers to machines that are fully autonomous, conscious, and therefore self-directing. AI is a core component of the AMs makeup, enabling them to function independently, make decisions, and adapt and operate in changing environments.

AM/MAMs will challenge our idea of consciousness, our values, and what we define as life. We need to be ahead of the challenge in both legal and ethical terms, create stability and confidence, and safeguard against overlooking any evolving rights and duties that will arise from AM development. This is especially vital in the military context, where the extra governance of International Humanitarian Law (IHL) will impact MAMs and the operational environment, due to the unique challenges and need for clear accountability.

Due to being a futuristic challenge, it is an area where the literature highlights differing views and that a general consensus is hard to find.[1] This is both in terms of the rights and duties AMs and MAMs could owe us, and if there will ever be a time when we have to recognise and uphold values, rights and duties to AMs/MAMs. Consequently, the literature is scant on agreeing what values, rights and duties for AMs/MAMs could look like, which this thesis stresses is a significant risk when looking at the development and deployment of MAMs. Nevertheless, there is an expectation that humans share certain universal values, which the

[1] For example, the literature highlights differing views as to whether to treat AMs as objects or agents, along with whether machines will ever be conscious. These, along with other challenges, are highlighted in the literature review.

United Nations stated as peace, freedom, social progress, equal rights, and human dignity[2], so it would appear a good start for AM/MAMs to understand and uphold these values.

The European Parliament (EP) commissioned a report into Civil Law Rules on Robotics, which highlighted that they are all too aware of the potential legal and ethical issues posed by the technological advances.[3] Our laws need to be future proof and with the decision to leave the EU[4], gaps could quickly start appearing. Supporting this view, Savirimuthu asserts that "there are considerable difficulties in ascribing responsibility through the formulation of clear and precise definitions of rights, duties and obligations in an environment where technological change seems to be relentless."[5] The ethico-legal gaps will have far reaching consequences for liability, the rights and duties we owe AMs, and the unique challenges MAMs will present. This is due to the development pace of the technology. Indeed, when the author started this research, ChatGPT did not exist, yet is now a tool widely used in day to day life.

It is this technological pace that makes future innovation hard to predict, with some concepts, such as machine consciousness, unimaginable to some people. Nonetheless, the literature demonstrates that machine consciousness is a real possibility within our lifetime

[2] United Nations, 'Universal values - peace, freedom, social progress, equal rights, human dignity - acutely needed, Secretary-General says at Tübingen University, Germany' (2003) United Nations Press Release < https://www.un.org/press/en/2003/sgsm9076.doc.htm > accessed 3 September 2024

[3] European Parliament, '*DRAFT REPORT with recommendations to the Commission on Civil Law Rules on Robotics'* (2015/2103(INL))

[4] The EU are investing heavily in research into autonomous vehicles, which the UK has benefited from, however the full impact of Brexit and the status of EU research, guidelines and safeguards within the UK is still not clear.

[5] Savirimuthu, Joseph, "Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence" Patrick Lin, Keith Abney and Ryan Jenkins (eds), International Journal of Law and Information Technology, Volume 26, Issue 4, Winter 2018, Pages 337–346, <https://doi.org/10.1093/ijlit/eay011> accessed 9 January 2023

due to the developing world of machine Deep Learning (DL), and this will have implications for the scope and validity of IHL with regards to MAMs.

DL is based upon deep artificial neural networks, which operate like a black box, making traceability, liability, and of vital importance accountability for IHL, very difficult to ascertain.[6] From the literature, it is not clear how we make decisions and what influences us, so trying to replicate this in another conscious entity is problematic and may prove too much of a risk within the military environment. Further, the research highlights that ethical theories of human free will, agency and autonomy have been pondered and debated only with humans and, in limited situations, animals, yet not in earnest for AMs and MAMs. Indeed, having explored the landscape, the research has shown that the ethico-legal facets have focused on humans exclusively having consciousness, personhood and autonomy, and not sharing these with any other entity, which is a significant gap. Nevertheless, we are embracing the technology along the pathway to machine consciousness (for example ChatGPT), yet developing and using AMs and MAMs will have ethical challenges, specifically with regards to IHL.

As a consequence, we will need to discuss decision-making, accountability, and specifically the uniqueness of these within the military environment, which must comply with IHL. Thus, this thesis focuses on the gaps in understanding decision-making and accountability in the context of conscious AMs and MAMs. It looks at where IHL needs to be further developed,

[6] Jai Galliott, Duncan MacIntosh, and Jens David Ohlin, *Lethal Autonomous Weapons. Re-Examining the Law and Ethics of Robotic* Welfare (2021 Oxford University Press)

values extended, and discussions encouraged, to recognise and accommodate MAMs, which is identified here as the true value alignment problem (TVAP).

1.1 Research Aims

This thesis has four research aims, which are:

1) To explore and understand the legal landscape of AI and IHL.
2) Exploration to understand if machine consciousness has been designed to work with IHL.
3) Examination into whether the current code of IHL ethics aligns with machine consciousness decision-making.
4) To understand whether MAMs truly align to the principles of IHL

1.2 Methodology

This is an area of law that traditionally follows a doctrinal approach.[7] Consideration was given by the author to both qualitive[8] and quantitative[9] approaches, however, when reviewing the

[7] Terry Hutchinson and Nigel Duncan, 'Defining and Describing What We Do: Doctrinal Legal Research' [2012] Deakin Law Review 17. 83-119

[8] UK Research and Innovation, 'What is social science?' (*UKNI, 31 March 2022*) https://www.ukri.org/who-we-are/esrc/what-is-social-science/qualitative-research/ > accessed 8 August 2024

[9] UK Research and Innovation, 'What is social science?' (*UKNI, 31 March 2022*) https://www.ukri.org/who-we-are/esrc/what-is-social-science/qualitative-research/ > accessed 8 August 2024

literature, it was seen that there was ample literature, which showed clear gaps that this thesis will address. It is the view here that a doctrinal approach is best placed for understanding the legal framework and ensuing a systematic and objective analysis of the law and relevant texts, which results in revealing where the gaps are. Further, it was considered that obtaining new data in this area would be difficult as conscious AMs are still a theorical concept and, at the time of writing, not in existence.

Therefore, the research aims will be answered through secondary research, technology articles, organisational reports, academic libraries, and educational institutions. Documentary analysis will focus on obtaining data from existing literature including black letter law, articles, official reports, grey literature (e.g., Ministry of Defence documentation), caselaw, newspapers and journals.

As this AM and MAM technology development is being led by technologists and technology companies, such as Microsoft[10], IBM[11], and Tesla[12], the author conducted significant research into technology companies and their innovation, reading their publications and case studies. This provided insight into how far away machine consciousness is, the level of ambition, and where the current technology is being deployed. The aided the literature search through providing insight into key developments in the area and searches that were needed.

The author identified the relevant literature to read through specific search terms and through the references cited in the resulting articles and books.

[10] Microsoft Research, 'Artificial Intelligence' (*Microsoft, 2025*) < https://www.microsoft.com/en-us/research/research-area/artificial-intelligence/? > accessed 4 January 2025

[11] IBM Research, 'What's Next in AI is foundation models at scale' (*IBM, 2025*) < https://research.ibm.com/artificial-intelligence > accessed 4 January 2025

[12] Tesla AI and Robotics, 'AI and Robotics' (*Tesla, 2025*) < https://www.tesla.com/en_gb/AI > accessed 4 January 2025

The search terms the author employed in the online research libraries via the UCLAN library and Google were: AI history and development; machine consciousness; artificial intelligence and machine consciousness; human consciousness; animal consciousness; military autonomous machines; ethics of war; EU and IHL laws on autonomous machines; British Rules of Engagement; British Army duty of care; rights and duties of humans in UK; Should machines be able to have rights and duties.

1.3 Structure of Thesis

The thesis is split into 7 chapters, of which this introduction is chapter 1. Chapter 2 is the literature review, which starts with the history and evolution of AMs, moving through to the present-day AMs and MAMs, and emphasises current day thinking on the ethico-legal risks and challenges. It highlights not only the value alignment problem (VAP), but also what the author terms the true value alignment problem (TVAP), which is a fundamental gap in the literature. Chapter 3 is focused on research aim 1 and explores and understands the legal landscape of Al and IHL, drawing attention to the TVAP. Chapter 4 is focused on research aim 2, which aims to understand if machine consciousness has been designed to work with IHL and the emergence of the TVAP. Chapter 5 is focused on research aim 3, which examines whether the current code of IHL ethics aligns with machine consciousness decision-making and the implications for the TVAP. Chapter 6 concentrates on research aim 4, which aims to understand and explore how MAMs align to the principles of IHL and centres on the TVAP. Finally, chapter 7 draws together the discussion from within the research questions and emphasises the impact and risks of the gaps in current research, including the IHL and the

VAP and the TVAP. The chapter makes recommendations for addressing the gaps and mitigating the risks. Further, the chapter highlights areas for possible future research.

2. Literature Review

2.1 Introduction

The literature review discusses unconscious AMs, unless stated otherwise. This is due to the lack of literature at the time of writing that considered conscious AMs. Further, although the literature review, in parts, is categorised by the research aims, it should be noted that all the literature had relevance in meeting the research aims, and in forming the discussion and recommendations.

In 1589, Queen Elizabeth refused a patent to William Lee for a knitting machine, because she thought it would put people out of work and she valued 'her' people foremost.[13] This trepidation around emergent technologies has remained throughout hundreds of years and is still just as evident today with AM innovation. In fact, machine consciousness has often been the storyline of many Sci-Fi Hollywood films, such as the 1984 movie Terminator, where machines are set to take over the world. Yet, whilst the risks to humanity are typically central to these movies, little thought is given to the rights of the machine. Our technology may have come a long way since William Lee's knitting machine, but our attitude to technology and putting human values and autonomy above a machine, has not.

The word 'autonomy' derives from the Greek words auto (self) and nomos (law). According to Beauchamp and Childress, autonomy is the ``personal rule of the self that is free from both controlling interferences by others and from personal limitations that prevent meaningful

[13] Christ's College University of Cambridge, 'William Lee' (Christ's College University of Cambridge Online) <https://www.christs.cam.ac.uk/william-lee> accessed 30 September 2022

choice."[14] Thus, an autonomous person is someone who `freely acts in accordance with a self-chosen plan'' as agreed by Kant.[15] Majeed provides a key descriptor of the controls we humans operate under, saying "ethics may be simply described as 'the intrinsic control of good behaviour'. This contrasts with 'law' that acts as the 'extrinsic control of good behaviour'."[16] Majeed's quote highlights the multiple dimensions of good versus bad, law versus ethics, values and morality, and is used as a golden thread throughout this thesis to draw together the framework AMs, specifically military AMs (MAMs) will need to navigate and operate within. Majeed's view is mirrored by Lijadi and the values she identifies that define 'a good life'[17], which encompass the UN's previously mentioned list[18].

In 1956, John McCarthy first introduced us to the term 'Artificial Intelligence (AI)'.[19] Over the years it has been used to describe machines that are able to carry out tasks that typically require human intelligence such as decision-making and speech recognition.[20] Since the 1950s, there have been significant technological developments, which have led to fully autonomous machines being developed for use by the military, by the medical profession and more recently the development of autonomous vehicles. Although semi-autonomous machines[21] have featured in manufacturing and mining industries for many years, these

[14] Beauchamp, T.L. & Childress, J.F, *Principles of biomedical ethics (4th ed)* (1994). (Oxford University Press.) Pg 37.

[15] Beauchamp, T.L. & Childress, J.F, *Principles of biomedical ethics (4th ed)* (1994). (Oxford University Press.) Pg 37.

[16] A.B.A Majeed, 'Roboethics - Making Sense of Ethical Conundrums' (2017) Procedia Computer Science, Volume 105, 2017 < https://doi.org/10.1016/j.procs.2017.01.227 > accessed 19 March 2019.

[17] Anastasia Aldelina Lijadi , 'What are universally accepted human values that define 'a good life'? Historical perspective of value theory' (2019) WP-19-006 IIASA < https://pure.iiasa.ac.at/id/eprint/16049/1/WP-19-006.pdf > accessed 4 September 2024

[18] United Nations, 'Universal values - peace, freedom, social progress, equal rights, human dignity - acutely needed, Secretary-General says at Tübingen University, Germany' (2003) United Nations Press Release < https://www.un.org/press/en/2003/sgsm9076.doc.htm > accessed 3 September 2024

[19] In 1956, John McCarthy organised a conference at Dartmouth Summer Research Project on Artificial Intelligence, where the term Artificial Intelligence (AI) was born.

[20] Michael Copland, 'What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?' (*Nvidia*, 29 July 2016) <https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/ > accessed on 24 April 2017

[21] From the literature read to date, the general consensus this this is Shared human and machine control.

machines are programmed to repeatedly perform a specific function under controlled conditions.[22] We are taking AMs out of the pre-defined, controlled conditions, that they are used to and moving them into the chaos of complex environments, specifically the military environment, which reveals novel legal and ethical questions and concerns. In 2014, Elon Musk tweeted, "we need to be super careful with AI. Potentially more dangerous than nukes, "[23] with Stephen Hawking's saying "success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks."[24] These proclamations are from two of the leading scientific and technological minds in this field at this time and it would be foolhardy to ignore them. They not only acknowledge the massive achievement and benefits AI will bring, but can foresee the risks and negative implications.

With this in mind, the problem of who is liable when an act or omission is exclusively the result of the AM/MAMs needs to be established, along with our rights and duties towards AM/MAMs. Our laws and ethics, including International Humanitarian Law (IHL), do not encompass AM/MAMs and does not consider artificial consciousness. Savirimuthu's statement of our "belief in the invisible hand in guiding development" stands as a warning that we must purposefully lead AM development, thus actively look out for, and manage, all the ethico-legal challenges that will arise. [25]

[22] Ugo Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts* (2013 edn, Springer)

[23] Twitter, @elonmusk on 7:33 pm - 2 Aug 2014.

[24] The Independent, 'Stephen Hawking: 'Transcendence looks at the implications of artificial intelligence - but are we taking AI seriously enough?" (The Independent, 01 May 2014 < https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html > accessed 10 November 2017.

[25] Savirimuthu, Joseph, "Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence" Patrick Lin, Keith Abney and Ryan Jenkins (eds), International Journal of Law and Information Technology, Volume 26, Issue 4, Winter 2018, Pages 337–346, <https://doi.org/10.1093/ijlit/eay011> accessed 9 January 2023

The literature review will now be discussed in line with the forthcoming research aims. The research aims will focus on criminal law, IHL, and the associated challenges and potential liability, yet at times will touch on civil law to provide background and context.

2.2 Research Aim 1: To explore and understand the legal landscape of AI and IHL

Our legal framework is extensive from both a civil and criminal perspective and must be reviewed in light of AMs development. To note, current English civil and criminal law do not specifically cover AMs.

2.2.1 Development of AI

Turner highlights the struggle of how we are to treat AMs, be that as a thing, an object or a person.[26] Solaiman[27] considers AMs to be property and thus 'objects' of law as opposed to 'subjects', a view shared by Leenes and Lucivero.[28] Leenes and Lucivero[29] argue that AMs are currently treated as tools and a human will remain legally responsible for the AM's actions, thus ensuring the AM operates within the confines of the law. In the American case of Stanley[30], it was professed that autonomy and self-determination are not deemed as grounds for bestowing legal rights on any entity, although this does not determine the UK view, with

[26] Jacob Turner, 'Robot Rules. Regulating Artificial Intelligence', 2019, Palgrave Macmillan.

[27] S.M Solaiman, 'Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy' (2017) Artif Intell Law 25, 155–179 (2017) <https://link.springer.com/content/pdf/10.1007/s10506-016-9192-3.pdf > accessed on 17th October 2018.

[28] Leenes R, Lucivero F (2014) Laws on robots, laws by robots, laws in robots: regulating robot behaviour by design. Law Innov Technol 6(2).

[29] Leenes R, Lucivero F (2014) Laws on robots, laws by robots, laws in robots: regulating robot behaviour by design. Law Innov Technol 6(2).

[30] Matter of Nonhuman Rights Project, Inc v Stanley [2015] NY Slip Op 31419(U).

the English legal view focused on product liability. Bryson[31] strongly argues against terming an AM a moral agent, as it simply serves as a method to repeal our own responsibility, which is view shared by Loh and Loh, although they only considered limited autonomous vehicles.[32] Sandberg[33] discusses the EU draft plan[34] of turning AMs into electronic persons and highlights that the EU's plan overlooks the opportunity to embed our moral and legal policies into the software of the AMs. He recognises that we already, "treat animals as between legal subjects and objects: they are objects, but it is recognised that they can do things their owners cannot foresee."[35] Critiquing the EU report, Sandberg[36] says that "most of the text deals with robots. Yet it is the software that makes them autonomous and problematic, and the software can reside in ill-defined places like the cloud, remote jurisdictions or the blockchain",[37] suggesting that stricter software liability will be required, and that AMs could be fitted with 'black boxes' that would record all their actions. This could then assist in identifying fault should an incident occur.

Brozek and Jakubiec do not feel AMs can be permitted to be legal agents and believe ideas as such will be confined to 'law in books' and not transferring to 'law in action', due to their opinion that AMs cannot be seen as originators of their own actions.[38]

Mills & Reeve paper on autonomous vehicles (AVs) and the law, stresses that the law and regulation normally advance through developments in public behaviour and developments in

[31] JJ Bryson, 'Robots should be slaves. In: Wilks Y (ed) Close engagements with artificial companions: key social, psychological, ethical and design issue' [2010] John Benjamins Publishing Company, Amsterdam.

[32]In: Partick Lin, Ryan Jenkins and Keith Abney, *Robot Ethics 2.0* (2020, Oxford University Press) 38-47.

[33] Anders Sandberg, 'Law-abiding robots?', (2016) University of Oxford < https://www.oxfordmartin.ox.ac.uk/opinion/view/340 > accessed on 24 October 2017.

[34] European Parliament, '*DRAFT REPORT with recommendations to the Commission on Civil Law Rules on Robotics'* (2015/2103(INL))

[35] Anders Sandberg, 'Law-abiding robots?', (2016) University of Oxford < https://www.oxfordmartin.ox.ac.uk/opinion/view/340 > accessed on 24 October 2017.

[36] Anders Sandberg, 'Law-abiding robots?', (2016) University of Oxford < https://www.oxfordmartin.ox.ac.uk/opinion/view/340 > accessed on 24 October 2017.

[37] Anders Sandberg, 'Law-abiding robots?', (2016) University of Oxford < https://www.oxfordmartin.ox.ac.uk/opinion/view/340 > accessed on 24 October 2017.

[38] Bartosz Broozek and Marek Jakubiec, 'On the legal responsibility of autonomous machines' [2017] Artif Intell Law (2017) 25:293-304.

what the public deem as acceptable.[39] They make an interesting observation that society tolerates, and is accepting of the fact, there will be deaths and injuries on the road, from which the driver can often walk away, however Mills & Reeve see no other area where society is so tolerant of death and injury brought about through human error.[40] It is here that they can see AVs having a significant positive benefit for society but are hasty to add that AVs will not be a golden goose. This is in stark contrast to MAMs where the public are suspicious and cautious of any 'benefits' as acknowledged in the DoD report.

2.2.2 Liability and Ownership

There is support for owner's liability, akin with animal owner's liability, as confirmed by Weng et al[41] and Leenes and Lucivero[42], whilst others like Bertolini[43] favour leaving AM liability to the manufacturers' covered by product liability and away from criminal liability. Arguably, assigning legal personality to AMs risks absolving humans of liability and, in doing so, reducing the effectiveness of deterrence.[44]

Some academics and lawyers see product liability as being the solution, and feel that it is sufficient to cover the issue of AM liability.[45] Supporting this is the DfT who point towards

[39] Mills & Reeves, 'Why we should get use to the idea that self-driving cars will sometimes crash' (2015) Mills & Reeve < https://www.mills-reeve.com/files/Uploads/Documents/Autonomous-Vehicles-Article-Is-it-an-ethical-or-a-legal%20question.pdf > accessed on 24 November 2017.

[40] Mills & Reeves, 'Why we should get use to the idea that self-driving cars will sometimes crash' (2015) Mills & Reeve < https://www.mills-reeve.com/files/Uploads/Documents/Autonomous-Vehicles-Article-Is-it-an-ethical-or-a-legal%20question.pdf > accessed on 24 November 2017.

[41] YH Weng, CH Chen and CT Sun CT, 'Toward the human–robot co-existence society: on safety intelligence for next generation robots' [2009] Int J Social Robot 1:267–282.

[42]R Leenes and F Lucivero, ' Laws on robots, laws by robots, laws in robots: regulating robot behaviour by design' [2014] Law Innov Technol 6(2):193–220.

[43] A Bertolini, 'Robots as products: the case for a realistic analysis of robotic applications and liability rules' [2013] Law Innov Technol 5(2):214–247.

[44] JJ Bryson, 'Robots should be slaves. In: Wilks Y (ed) Close engagements with artificial companions: key social, psychological, ethical and design issue' [2010] John Benjamins Publishing Company, Amsterdam.

[45] Neil M Richards and William D Smart, 'How Should the Law Think About Robots?' (10 May 2013) < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2263363> accessed on 17th November 2017;

product liability and safety law to address the emergence of AVs. In discussing AVs, Schelleken[46] thinks manufacturers will be able to avoid liability under the current legislation by showing, "that the state of scientific and technical knowledge at the time when he put the product into circulation was not such as to enable the existence of the defeat to be discovered,"[47] although this defence is equally as applicable to manufacturers of all other AMs.

Despite supporters of product liability legislation being the legal answer to liability, the DfT stress the need for greater legal certainty, as they see gaps in the current legislation and where liability lies, which is problematic for all and could stifle innovation.[48] This is supported by both Pagallo[49] and Schellekens[50] who warn that the level of liability and risk the manufacturers are exposed to will affect the launching of AMs including fully AVs. Mills and Reeve[51] believe the lack of legal certainty is hindering AV innovation, with vital questions needing answers by the legislators. They deem an international treaty essential for reassurance and conformity.

Pagallo[52] believes that relying on the umbrella of product liability seems inadequate, and grouping the household iron together with a self-thinking machine seems foolhardy. Nevertheless, Owen[53] points the finger of liability in the direction of the manufacturer stating, "it is the human actor who programmes the computer, and only the human actor who is

R Leenes and F Lucivero, 'Laws on robots, laws by robots, laws in robots: regulating robot behaviour by design' [2014] Law Innov Technol 6(2):193–220).

[46] Maurice Schellekens, '*Self-driving cars and the chilling effect of liability law*', [2015] Computer Law & Security Review, Volume 31, Issue 4, August 2015, Pages 506–517.

[47] Maurice Schellekens, '*Self-driving cars and the chilling effect of liability law*', [2015] Computer Law & Security Review, Volume 31, Issue 4, August 2015, Pages 506–517.

[48] DfT, 'The Pathway to Driverless Cars: Summary Report and Action Plan' (2015) < https://www.gov.uk/government/publications/driverless-cars-in-the-uk-a-regulatory-review > accessed on 5 November 2016.

[49] Ugo Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts* (2013 edn, Springer).

[50] Maurice Schellekens, '*Self-driving cars and the chilling effect of liability law*', [2015] Computer Law & Security Review, Volume 31, Issue 4, August 2015, Pages 506–517.

[51] Mills and Reeve, 'Legalising autonomous vehicles: Why it's a drug related question' (2016) Mills-Reeve.com < https://www.mills-reeve.com/legalising-autonomous-vehicles/ > accessed 2 November 2011.

[52] Ugo Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts* (2013 edn, Springer) .

[53] Tim Owen, Crime, Genes, Neuroscience and Cyberspace (Palgrave Macmillan 2017).

capable of formulating and acting upon decisions."[54] which Brown[55] counters by arguing machines will develop beyond our influence and control. Asaro[56] also challenges Owen's stance by saying "autonomous artificial agents can act in the world independently of their designers and operators"[57] and stresses they will act upon their own decisions, with unpredictable and unintended actions and effects. This is supported by Franklin and Graesser[58] who also advise of formalising the definition of an autonomous agent to plainly distinguish a software agent from any other software program.

Asaro[59] questions how people, including designers and manufacturers, will remain liable and take moral accountability for AMs when the foreseeability of risk is currently indeterminable (e.g., learns behaviour not foreseeable by the designer). Like Pagallo,[60] Asaro draws parallels between the liability for young children and AMs. In a similar vein, Duffy and Hopkins[61] pose the idea of treating AVs like dogs when it comes to the question of liability, thus the owner remains liable with the standard being that of strict liability, which is also favoured by the EU[62]. Whilst this may seem acceptable for today's AMs, especially by manufacturers and owners, Asaro[63] questions the suitability for future AMs, which would be far more advance.

[54] Tim Owen and Julie Owen, 'Virtual Criminology: Insights from genetic-social science and Heidegger' [2015] Journal of Theoretical and Philosophical Criminology 7.17-31.

[55] Sheila Brown, 'The criminology of hybrids: Rethinking crime and law in technosocial networks' (2006) *Theoretical Criminology*, *10*(2), 223–244 < https://doi.org/10.1177/1362480606063140> accessed on 17 November 2017.

[56] Peter Asaro, 'The Liability Problem for Autonomous Artificial Agents' [2016] Association for the Advancement of Artificial Intelligence.

[57] Peter Asaro, 'The Liability Problem for Autonomous Artificial Agents' [2016] Association for the Advancement of Artificial Intelligence.

[58] S. Franklin and A Graesser, 'Is it an agent or just a program? a taxonomy for autonomous agents' [1997] Müller J.P., Wooldridge M.J., Jennings N.R. (eds) Intelligent Agents III Agent Theories, Architectures, and Languages. ATAL 1996. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 1193. Springer, Berlin, Heidelberg.

[59] Peter Asaro, 'The Liability Problem for Autonomous Artificial Agents' [2016] Association for the Advancement of Artificial Intelligence.

[60] Ugo Pagallo, The Laws of Robots: Crimes, Contracts, and Torts (2013 edn, Springer).

[61] Sophia Duffy and Jamie Hopkins, 'Sit, Stay, Drive: The Future Of Autonomous Car Liability" (2014) < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2379697 > accessed on 10 December 2018.

[62] European Parliament, '*DRAFT REPORT with recommendations to the Commission on Civil Law Rules on Robotics'* (2015/2103(INL)).

[63] Peter Asaro, 'The Liability Problem for Autonomous Artificial Agents' [2016] Association for the Advancement of Artificial Intelligence.

Caliskan et al contend that we are teaching AMs to be prejudiced and AMs are not equipped to wilfully counteract their learnt biases.[64]

Whilst manufacturers may accept a moderate degree of product liability for AMs, for example, for the quality of the components they are made from, it is highly likely that manufacturers of AMs will not so keen to be held criminally liable for a machine that self-learns based on stimuli and its environment, all of which are outside the control of the manufacturer. This becomes even more likely when looking at MAMs, who will be deployed and controlled way beyond the manufacturers scope and span of control. Manufacturers actively look to limit or remove their responsibility wherever possible, especially under Acts as severe as the Corporate Manslaughter and Corporate Homicide Act 2007 (CMCHA). Holder et al[65] recognise that with the dawn of AMs, where the criminal elements of mens rea and actus reus may be harder to define, along with apportioning culpability. They argue that product liability may be inflexible or too narrow and therefore becomes questionable how well the current legislation will cope.

2.2.3 Criminal Liability

As stated above, criminal law has two elements that must be proved for there to be criminal liability; The first is a criminal act (actus reus), and the second a criminal mind (mens rea). If either one of these elements are missing, then no criminal responsibility can be enforced, as

[64] Aylin Caliskan, Joanna J Bryson and Arvind Narayanan, 'Semantics derived automatically from language corpora contain human-like biases' (2017) Science356,183-186 https://www.science.org/doi/10.1126/science.aal4230 accessed on 9 August 2020.

[65] Chris Holder, Vikram Khurana, Faye Harrison, and Louisa Jacobs, 'Robotics and law: Key legal and regulatory implications of the robotics age (Part I of II)' [2016] Computer Law & Security Review, Volume 32, Issue 3.

illustrated by Hallevy[66] and his parrot analogy; "A parrot is capable of repeating words it hears, but it is incapable of formulating the mens rea requirement…"[67] Criminal law is explored in detail in Research Question 1 to understand the challenges we may face in ascribing liability to a human who is not deciding or directing a AM/MAMs actions. Criminal law has the highest evidential bar and can be applied to military personal acting outside of IHL.

Pagallo[68] asks how, if at all, the mens rea can be established for an AM/MAM and if it would be just to hold them criminally accountable. He suggests that if an AM/MAM 'decides' to take an action that leads to the harm or even death of a human, for example a military machine or an AV, then it could have the required mens rea for criminal liability, however Pagallo[69] does not provide any firm conclusions and merely poses the questions with discussion around the implications. Hallevy[70] has dedicated much attention to the subject of AM criminal responsibility and, at the time of writing, is the only person who proposes three models that are formed around AMs and not simply just trying to bend existing laws to fit. His three models are; 1) the perpetration-by-another responsibility model, 2) the natural-probable-consequence responsibility model, and 3) the direct responsibility model.[71] Hallevy[72] views these three models as a solution to the problem of AMs meeting the actus reus and mens rea

[66] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

[67] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

[68] Ugo Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts* (2013 edn, Springer).

[69] Ugo Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts* (2013 edn, Springer).

[70] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

[71] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

[72] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

requirements for criminal responsibility and are supported by Scholten.[73] Solaiman[74] and Charney[75] offer support to the theory of the models do not yet believe they can be implemented.

2.2.4 Military Liability

When looking at the military setting, one must understand the unique laws and frameworks surrounding it. International Humanitarian Law (IHL) also known as the 'laws of war' or the Laws of Armed Conflict (LOAC), is the legal framework relevant to armed conflict and occupation, of which all UN member states abide by, including for the development of autonomous weapon systems (AWS), which are used in armed conflict. IHL has a foundation in JWT and is underpinned by the principles of distinction, necessity, proportionality, and humanity. It contains the duty to carry out legal reviews of the development, procurement, and implementation of new weapons, as specified by Article 36 of Additional Protocol I to the Geneva Conventions. Currently all UK military AWS and MAMs are under human control, and it is a human who takes accountability. At the time of writing, and to the best of the authors knowledge, no plans exist to use fully autonomous military machines or conscious MAMs within the military context; They will be used as tools to follow specific action as commanded by a human.

[73] Nina Scholten, 'The Robo-Criminal' [2019] Artificial Intelligence & Law (Fastcase) 263.

[74] S.M Solaiman, 'Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy' (2017) Artif Intell Law 25, 155–179 (2017) <https://link.springer.com/content/pdf/10.1007/s10506-016-9192-3.pdf > accessed on 17th October 2018.

[75] Rachel Charney, 'Can Androids Plead Automatism? A Review of When Robots Kill: Artificial Intelligence Under the Criminal Law by Gabriel Hallevy' (2015) 73 U T Fac L Rev 69 at 70.

The British Army (BA)[76] has established their 'Rules of Engagement' (ROE), which are the internal military rules that define the conditions, circumstances, degree, and the way that force may be used. The BA ROE are in addition to the protocol in the Geneva convention, and adds in further safeguarding rules, such as establishing a clear framework for decision making under 'command and control', incorporating additional legal and ethical policies, and setting a risk management framework in order to reduce the likelihood of non-combatant harm. However, the BA ROE are contained within JSP 389, which is a restricted document and therefore not discussed in detail further. The Armed Forces Act (AFA) 2006, s42-49, addresses criminal conduct of Service personnel, which is deemed as acts outside of those authorised by the BA, for example, murder and rape, as highlighted in the Marine A case.[77] In fact, MAMs could actually reduce the harm caused by war and cases such as Marine A, due to not being driven by emotion, as their lack of emotions will ensure their judgement is not blurred. Indeed, "'fear and hysteria are always latent in combat, often real, and they press us toward fearful measures and criminal behaviour.' Autonomous agents need not suffer similarly."[78] Adams[79] warns that "military systems (including weapons) now on the horizon will be too fast, too small, too numerous and will create an environment too complex for humans to direct."[80] This again highlights the need for clarity of legal liability and responsibility, as the risks and consequences are even greater and could be grave. Specifically, lack of a clear legal framework risks issues regarding jurisdiction and the legal framework the State/weapon is

[76] British Army is distinct from the Royal Navy and the Royal Airforce. All tasks in the British Army are directed by a soldier: https://www.army.mod.uk/who-we-are/.

[77] Discussed in Chapter 5.

[78] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

[79] T Adams, "Future Warfare and the Decline of Human Decision-making" [2002] Parameters, US Army War College Quarterly, Winter 2001-02.

[80] T Adams, "Future Warfare and the Decline of Human Decision-making" [2002] Parameters, US Army War College Quarterly, Winter 2001-02.

operating under, e.g., Geneva Convention. Further, although out of scope of this thesis, issues arise regarding the jurisdictional regulation of private companies, who will be part of the development and procurement processes, and may be governed by the headquarters of the country, rather than the where the MAM is deployed.

As long as the British Army (BA) adhere to International Humanitarian Law (IHL), then their actions are legitimate, however MAMs may change this. Indeed, Cook and Syse[81] emphasise the challenge for law of keeping abreast with military technology and practices. They see the law as needing "reinterpretation" to stay "relevant and useful in guiding that changing activity."[82] This is certainly true with the use of MAMs, especially conscious MAMs.

2.2.5 Protection for MAMs

Professor Marco Sassòli[83] sees many advantages to MAMs, specifically, that he believes they will not hate or be afraid, but it is the view here that this must not be exploited by humans, for example, deliberately exposed and/or trained on bias or out of date data, to undertake actions stemming from individual viewpoints/hidden agendas. Protections must be put in place, which could require an independent process similar to the car MOT process, which validates and checks how MAMs are behaving and allows them share concerns. Whilst of interest, this type of process is outside of the scope of this thesis.

[81] Martin L. Cook & Henrik Syse, "What Should We Mean by 'Military Ethics'?" [2010] Journal of Military Ethics, 9:2, 119-122.

[82] Martin L. Cook & Henrik Syse, "What Should We Mean by 'Military Ethics'?" [2010] Journal of Military Ethics, 9:2, 119-122.

[83] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

In addition to State responsibilities, the Ministry of Defence (MoD)[84] further owes a duty of care towards its soldiers, as found in the case of Smith.[85] In Smith, it was claimed that the MoD breached its obligation to safeguard life protected by ECHR art 2.[86] This was despite the MoD arguing that the claims under ECHR art 1[87] should be struck out. In fact, the Chilcot report was issued in 2016, which clearly showed that the MoD knew about the vulnerabilities that led to the deaths yet failed to take action. Although this is a civil case, it is used in this thesis to highlight the responsibility the BA owes to soldiers. The case additionally raised the issue of the civilian courts reviewing with decisions of the battlefield, yet without the military training, expertise, or context, which was stressed in the Constitution Committees report[88] and emphasised by Morgan[89]. This duty of care has not been established for MAMs.

The literature fails to acknowledge, or even consider in passing, that the MoD may have a future duty of care towards MAMs, especially conscious ones. The thinking and research around deploying MAMs is limited and tends to be confined to sci-fi films. Nevertheless, it is vital to consider that conscious MAMs will have their own goals and free will akin to a human soldier, yet they will not be afforded the luxury of choosing their own path in life, thus will be designed with a purpose, conditioned and deployed as humans see fit. For the author, this sits uncomfortably, and it is the view here that there needs to be a framework for the duty of care for MAMs to be developed and implemented before they are deployed. Therefore,

[84] The MoD is a ministerial department that oversees the BA.

[85] Smith and others (Appellants) v The Ministry of Defence (Respondent) Ellis (Respondent) v The Ministry of Defence (Appellant) Allbutt and others (Respondents) v The Ministry of Defence (Appellant) [2013] UKSC 41.

[86] Article 2 regards the right to life and provides that the State should safeguard life and take measures to investigate a death. Smith confirmed that the right to life secures the responsibility of the British government for the deaths of soldiers in combat, killed by enemy troops or illness, if their death is due to inadequate equipment or medical provisions/care. If forces serving abroad are not within the State's jurisdiction under Art 1 then the duties under Art 2 do not apply.

[87] Article 1 of the European Convention on Human Rights provides that rights and freedoms should be available to all those within the State's jurisdiction.

[88] House of Lords Constitution Committee, "Constitutional arrangements for the use of armed force" Second Report, 2013-4 Session.

[89] Dr. Jonathan Morgan, 'Military Negligence: Reforming Tort Liability after Smith v. Ministry of Defence' House of Commons Defence Select Committee (2013) https://www.biicl.org/files/6759_military_negligence_paper-_jonathan_morgan.pdf > accessed 19 October 2021.

research aim 1 will discuss the rights and duties for MAMs, and the legislative shifts around this.

2.3 Research Aim 2: Exploration to understand if machine consciousness has been designed to work with IHL

2.3.1 Understanding Machine Consciousness

Copland shows the progression of AI and explains the difference between AI, machine learning (ML) and deep learning (DL), discussing the potential that neural networks offers with regards to autonomy.[90] ML enables the machine to execute tasks and automate manual processes from human input that could be in text or audio form (e.g., Amazon's Alexa). DL is a wider part of ML and encompasses Artificial Neural Networks (ANNs). Cognitive computing advances self-learning machines further and was used in IBM's Watson project. The algorithms behind ML are often highly commercially sensitive and secretive, forming part of an organisation's intellectual property as exampled by Google and IBM[91], so understanding how they 'learn' is surrounded in commercial sensitivity and technical complexity[92]. To summarise, DL, ML, and AI are often described as a set of Russian dolls; DL is a subset of ML;

[90] Michael Copland, 'What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?' (*Nvidia*, 29 July 2016) <https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/ > accessed on 24 April 2017

[91] IBM Website, 'What is machine learning?' (IBM.COM) <https://www.ibm.com/topics/machine-learning?mhsrc=ibmsearch_a&mhq=what%20is%20machine%20learning > accessed 19 October 2020

[92] N Shea, and C D Frith, 'Dual-process theories and consciousness: the case for "Type Zero" cognition' [2016] Neurosci. Consci. 2016:niw005. doi: 10.1093/nc/niw005;
C M Signorelli, and X D Arsiwalla, 'Moral Dilemmas for Artificial Intelligence: a position paper on an application of Compositional Quantum Cognition' [2018] Quantum Interaction. QI 2018. Lecture Notes in Computer Science (Nice).

ML is a subset of AI, that is an umbrella term for any smart computer program. Thus, all ML is AI, but not all AI is ML.[93]

Hardesty[94] echoes the complexities of DL centring around algorithms (ANNs), which originate from the complex structure and functions of our brain. Dean highlights that the scalability of neural networks significantly enhances with the more data available and larger models.[95] Thus, increasing the ability to make an informed, 'meaningful choice'. Arthur Samuel views machine learning as the ability for machines to learning without being programmed, thus self-learning[96]. Indeed, a White Paper produced by the Symbiotic Autonomous Systems (SAS) Initiative[97] concludes that it is possible that machines will evolve awareness over the next decades and by 2050, consequently, AMs will be agents in their own right, which is supported by Global Neuronal Workspace (GNW) theory[98]. Matarić[99] defines an AM as acting "on the basis of its own decisions and is not controlled by a human."[100] It is the "not controlled by a human" that demonstrates some form of self-decision-making. Castro-Gonzalez[101] et al build on Matarić[102] by saying an AM acts on its own decisions, "in order to fulfill its goals" [103] and

[93] Kiltonsway, 'Artificial Intelligence (AI) vs. Machine Learning vs. Deep Learning' (Kiltonsway, 30 June 2021) < https://kiltonsway.mystrikingly.com/blog/artificial-intelligence-ai-vs-machine-learning-vs-deep-learning >accessed 13 June 2022.

[94] Larry Hardesty, 'Making computers explain themselves' (*MIT News*, 27 October 2016) <http://news.mit.edu/2016/making-computers-explain-themselves-machine-learning-1028 > accessed on 24 April 2017

[95] Google Research People, 'Jeffrey Dean' (Google Research) <https://research.google/people/jeff/ > accessed on 10 October 2020

[96] M. AlDarwish, "Machine Learning," (Carnegie Mellon) http://www.contrib.andrew.cmu.edu/~mndarwis/ML.html. > accessed 19 October 2017

[97] Roberto Saracco, Raj Madhavan, S. Mason Dambrot, Derrick de Kerchove, and Tom Coughlin, 'Symbiotic Autonomous Systems An FDC Initiative, White Paper' (2017) IEEE.org <https://digitalreality.ieee.org/images/files/pdf/sas-white-paper-final-nov12-2017.pdf > accessed 13 April 2019

[98] The global neuronal workspace model predicts that conscious presence is a nonlinear function of stimulus salience; i.e., a gradual increase in stimulus visibility should be accompanied by a sudden transition of the neuronal workspace into a corresponding activity pattern (Dehaene et al. 2003).

[99] Maja J Mataric , *The Robotics Primer*, (The MIT Press Cambridge, Massachusetts London, England 2007).

[100] Maja J Mataric , *The Robotics Primer*, (The MIT Press Cambridge, Massachusetts London, England 2007) Pg 2.

[101] Álvaro Castro-González, María Malfaz, J. F. Gorostiza, and Miguel A. Salichs, 'Learning Behaviours by an Autonomous Social Robot with Motivations' (2014), Cybernetics and Systems Vol. 45 Iss. 7,2014

[102] Maja J Mataric , *The Robotics Primer*, (The MIT Press Cambridge, Massachusetts London, England 2007)

[103] Álvaro Castro-González, María Malfaz, J. F. Gorostiza, and Miguel A. Salichs, 'Learning Behaviours by an Autonomous Social Robot with Motivations' (2014), Cybernetics and Systems Vol. 45 Iss. 7,2014

must have the action required for every situation or, if not equipped with this, learn how to interpret a situation and respond with the appropriate action. Consequently, the AM must know what action to execute in each situation. In the case that a robot does not have this knowledge, it must learn this relationship between situations and actions."[104] This could be said of humans as we have to learn the connection between situations and actions too. However, this interpretation is refuted by Brozek and Jakubiec[105] who say that all actions of an AM are traceable back to a human and is supported by Smart and Richard.[106] Supporting this, Wallach[107] believes designers should program the AMs for each situation, although avoids offering a solution to novel situations. The conscious AMs that are the focus of this thesis, will act independently without human oversight or control, akin to an adult human. Ignoring this problem will create challenges and risks later, especially for IHL where MAMs are currently viewed as weapons for use by a human. Further, Wallach[108] does not seem to comprehend the multitude of situations and scenarios AM will need to be programmed for, especially MAMs, which would quite frankly be impossible when considering the complexities of the battlefield and the stringent requirements of IHL. Brozek and Jakubiec[109] do not see AMs as originators of their own actions and see 'folk-psychology' as facilitating our perception of agency, for which they include emotions and the ability to feel pleasure and pain, and they

[104] Álvaro Castro-González, María Malfaz, J. F. Gorostiza, and Miguel A. Salichs, 'Learning Behaviours by an Autonomous Social Robot with Motivations' (2014), Cybernetics and Systems Vol. 45 Iss. 7,2014

[105] Bartosz Broozek and Marek Jakubiec, 'On the legal responsibility of autonomous machines' [2017] Artif Intell Law (2017) 25:293-304

106 Richards, Neil M. and Smart, William D, 'How Should the Law Think About Robots?' (10 May 2013). < https://ssrn.com/abstract=2263363>

[107] Wendell Wallach and Colin Allen, Moral Machines: Teaching Robots Right from Wrong: Teaching Robots Right from Wrong (2008) Oxford University Press

[108] Wendell Wallach and Colin Allen, Moral Machines: Teaching Robots Right from Wrong: Teaching Robots Right from Wrong (2008) Oxford University Press

[109] Bartosz Broozek and Marek Jakubiec, 'On the legal responsibility of autonomous machines' [2017] Artif Intell Law (2017) 25:293-304

do not see these as facets of AMs. Nevertheless, Sheridan[110], Hendriks[111] and Darling[112] emphasise the link between AM consciousness, the increase of social robots (e.g., Sophia) and human-robot social interaction, which will aid a shift in our perception of agency.

2.3.2 Risk and Critique of IHL

Consequently, to understand the risks AM/MAMs pose, we must understand the level of autonomy an AM/MAM has, and the level of control humans will retain, with special attention to accountability under IHL. When grading machine autonomy, Jha[113] is an advocate of a 10 level scale, with level 10 being a machine that can decide "everything and acts autonomously, ignoring the human operator completely"[114], as has full autonomy and agency. This is an informative framework and shares similarities with the SAE[115] scale, which has also been adopted by the US Department for Transport.[116] Nevertheless, the US DoD report on autonomy concluded that levels of autonomy are not particularly useful and often counter-productive as the attention tends to be on the hardware and software rather than focusing on the partnership of computer and operator working together to complete the actions and goals.[117]

[110] T B Sheridan, T. B, 'Human-robot interaction: status and challenges' [2016] Hum. Factors 58, 525–532. doi: 10.1177/0018720816644364.
[111] B Hendriks, B Meerbeek, S Boess, S Pauws, and M Sonneveld, 'Robot vacuum cleaner personality and behavior' [2011] Int. J. Soc. Robots 3, 187–195. doi: 10.1007/s12369-010-0084-5.
[112] K Darling, 'Extending legal protection to social robots: the effects of anthropomorphism, empathy, and violent behavior towards robotic objects' (2012) We Robot Conference 2012, April 23, 2012, < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2044797 > accessed 10 November 2017.
[113] Dr U C Jha, *Killer Robots: Lethal Autonomous Weapon Systems – Legal, Ethical and Moral Challenges*, (VIJ Books (India) 2016).
[114] Dr U C Jha, *Killer Robots: Lethal Autonomous Weapon Systems – Legal, Ethical and Moral Challenges*, (VIJ Books (India) 2016).
[115] Society of Automotive Engineers.
[116] Bill Canis, 'Issues in Autonomous Vehicle Deployment' (2017). Congressional Research Service 7-5700 < https://fas.org/sgp/crs/misc/R44940.pdf> *accessed on 30 October 2017*.
[117] Department of Defense Task Force Report: The Role of Autonomy in DoD System (2012) Department of Defense < https://fas.org/irp/agency/dod/dsb/autonomy.pdf > accessed on 5 November 2017.

To stress, it is the advancement of machines making independent decisions through the dawn of DL that makes this a growing problem of today and one we have to prepare and plan for. This is in stark contrast to the machines of yesterday, which took their commands from a human and acted as instructed or within set parameters[118], and all risks bourn by the human operator. Thus, it is the view here that the current scales of machine autonomy are limited, as they view a human as still maintaining ultimate control and do not extend to conscious AMs. Whilst the literature highlights the pros to humans remaining in control as accountability, public trust, and adherence to IHL, and the pros being human error, inconsistent and/or slower decision making, and protecting the harm to human soldiers, these are not all shared by the author. Indeed, the author argues throughout the research aims, that the pros identified in the literature, could return greater benefits if a MAM was in control.

The problems develop further, and therefore need an increased level of scrutiny, when deploying AMs in environments that are significantly challenging for humans. Indeed, a high risk, sensitive and politically dangerous environment to deploy AMs into is the military environments. This sensitivity and official secrecy of the algorithms is paramount with MAMs, whilst their deployment into the battlefield raises significant unique risks for IHL, especially around decision making, and consequently accountability. The IHL principles[119] as they current stand, focus on humans and thus do not accommodate machine consciousness. The regulation of MAMs in the military environment is complex due to the human terms on which it operates, including IHL, and because there is a requirement to dehumanise and further, to

[118] Some semi-autonomous machines, for example vehicles, will be given an objective/goal, but the method of meeting that objective is their choice, e.g. a vehicle driving from A to B and deciding when to brake, speed and lane assist.
[119] The principles are: necessity, humanity, proportionality, and distinction

‘de-machinise’ to meet the needs of the military ideology. Nevertheless, this is not a reason to apply more weighting the pros of human control mentioned above, or ignore the cons. As a consequence, research aim 2 is going to discuss the IHL considerations for conscious MAMs, the unique challenges posed by MAMs, and what the author identifies as the true value alignment problem (TVAP).

2.4 Research Aim 3: Examination into whether the current code of IHL ethics aligns with machine consciousness decision-making

2.4.1 Free Will of Decision Making

Luck and d’Inverno[120] identify the fundamentals for an ethical theory of agency and autonomy in software, specifying the connection between them. They classify programs as being an autonomous agent if they can operate independently of the program users and can modify their goals according to changes in circumstance. This involves free will and is supported by Chella and Manzotti.[121] Ishida and Chiba[122] and Harris[123] all endorse the view that our belief of ‘human’ free will derives from our conviction that we can act otherwise from what can be expected.

[120] Michael Luck and Mark d’Inverno, ‘A Formal Framework for Agency and Autonomy’ (1995) aaai.org < www.aaai.org/Papers/ICMAS/1995/ICMAS95-034.pdf> accessed on 20 October 2017

[121] Antonio Chella and Riccardo Manzotti, ‘Machine Free Will: Is free will a necessary ingredient of machine consciousness?’ (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017

[122] Yoshiteru Ishida & Ryunosuke Chiba “Free Will and Turing Test with Multiple Agents: An Example of Chatbot Design” [2017] Procedia Computer Science. 112. 2506-2518.

[123] John Harris, ‘Wonderwoman and Superman,’ (1992) Oxford

Chella and Manzotti[124] propose that self-determination is vital to free will and the ability to fulfil own goals, along with exploring compatibilism as a model for understanding machine free will. They question our understanding of free will and if humans really do have ultimate control over our behaviour, highlighting the fact human behaviour is influenced by many causes that are largely unknown. They argue that just because the source of an AMs behaviour originates from a human designer, that it does not automatically make the human any freer, and state, "a free choice is not a random one. A free choice is an act that is at the same time linked with the agent's past and unconstrained by external causes."[125] Yet this is disputed by Fuchs[126] who see humans as operating in environments imposed rather than chosen, thus not unconstrained. Chella and Manzotti[127] emphasise the difference between human automatic responses and conscious ones, drawing a parallel between the differences in automatic and free action, citing that, "the more a behaviour becomes automated (because of training or repetition) and the more it fades from consciousness thereby being subtract from conscious and free control",[128] which has been proven through research into cognitive control, such as the research by Snyder et al.[129] Chella and Manzotti[130] conclude by saying compatibilism is the only philosophical model for machine free will and will be based on the relationship between machine consciousness and freedom, "the two being the result of a

[124] Antonio Chella and Riccardo Manzotti, 'Machine Free Will: Is free will a necessary ingredient of machine consciousness?' (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017

[125] Antonio Chella and Riccardo Manzotti, 'Machine Free Will: Is free will a necessary ingredient of machine consciousness?'pg5, (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017

[126] Stephan Fuchs, 'Beyond Agency', [2001] University of Virginia

[127] Antonio Chella and Riccardo Manzotti, 'Machine Free Will: Is free will a necessary ingredient of machine consciousness?' (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017

[128] Antonio Chella and Riccardo Manzotti, 'Machine Free Will: Is free will a necessary ingredient of machine consciousness?'pg12, (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017

[129] Kristy M. Snyder & Yuki Ashitaka & Hiroyuki Shimada & Jana E. Ulrich & Gordon D. Logan, 'What skilled typists don't know about the QWERTY keyboard' (2013) Psychonomic Society, Inc. < https://link.springer.com/article/10.3758/s13414-013-0548-4 > accessed on 1 November 2017

[130] Antonio Chella and Riccardo Manzotti, 'Machine Free Will: Is free will a necessary ingredient of machine consciousness?' (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017

more efficient way to situate an agent in an unpredictable and largely unknown environment."[131]

Farnsworth[132] offers both a definition and fundamental requirements an agent must process to be considered as having free will, which complements Chella and Manzotti's[133] ideas and also distils free will down to an equation. Hooker[134] explains that "an agent is a being that is capable of action, and action is the exercise of agency."[135] In illustrating this, Hooker discusses the action of a mosquito biting a person and observes that a mosquito cannot be an agent because its action was just a result of chemistry and biology and not the result of any consideration on the mosquitos' part.[136]

2.4.2 Critique of Free Will

As with Chella and Manzotti,[137] Farnsworth[138] finds that the main obstacle to AM free will is the absence of being a Kantian whole, although Hooker[139] actively avoids Kantian language as he believes it implies ideas that are irrelevant to this particular argument. Unlike Chella and Manzotti,[140] who are suggestive that consciousness may not be linked to free will,

[131] Antonio Chella and Riccardo Manzotti, 'Machine Free Will: Is free will a necessary ingredient of machine consciousness?' pg12, (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017

[132] Keith Farnsworth, 'Can a Robot Have Free Will?' (2017) Entropy MDPI < https://pure.qub.ac.uk/portal/files/130713217/entropy_19_00237_v3.pdf > accessed on 1 November 2017

[133] Antonio Chella and Riccardo Manzotti, 'Machine Free Will: Is free will a necessary ingredient of machine consciousness?' (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017

[134] J Hooker, 'Autonomous Machines Are the Best Kind, Because They Are Ethical', (2016), Carnegie Mellon University < http://public.teppheaer.cmu.edu/jnh/agencyPost2.pdf > Accessed on 14 December 2017

[135] J Hooker, 'Autonomous Machines Are the Best Kind, Because They Are Ethical', (2016), Carnegie Mellon University < http://public.tepper.cmu.edu/jnh/agencyPost2.pdf > Accessed on 14 December 2017

[136] J Hooker, 'Autonomous Machines Are the Best Kind, Because They Are Ethical', (2016), pg2, Carnegie Mellon University < http://public.tepper.cmu.edu/jnh/agencyPost2.pdf > Accessed on 14 December 2017

[137] Antonio Chella and Riccardo Manzotti, 'Machine Free Will: Is free will a necessary ingredient of machine consciousness?' (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017

[138] Keith Farnsworth, 'Can a Robot Have Free Will?' (2017) Entropy MDPI < https://pure.qub.ac.uk/portal/files/130713217/entropy_19_00237_v3.pdf > accessed on 1 November 2017

[139] J Hooker, 'Autonomous Machines Are the Best Kind, Because They Are Ethical', (2016), Carnegie Mellon University < http://public.tepper.cmu.edu/jnh/agencyPost2.pdf > Accessed on 14 December 2017.

[140] Antonio Chella and Riccardo Manzotti, 'Machine Free Will: Is free will a necessary ingredient of machine consciousness?' (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017.

Farnsworth[141] is bold in saying that consciousness does not appear to be a requirement and, "the minimum complexity for a free-will system may be quite low and include relatively simple life-forms that are at least able to learn."[142] In contrast to Hooker, Farnsworth[143] sees his definition of free will as currently exclusive to living things, but does see the potential for it to be extended to AMs as the technology advances, which is a view supported by Basl.[144]

The debate has evolved to consider AMs, and if AMs could have or already have free will. Indeed, Owen and Owen[145] emphasise Heidegger's[146] 'Dasein' meaning, as, "to exist with a certain past, a personal socio-cultural history in the world and to have available 'ways to be' – a series of possibilities to exercise agency".[147] They then stress that, "this mineness' cannot be attributed to cyber technology",[148] as they strongly argue that, "no computer can be described as a "who" that is shaped and formed by existence in time, a creature with a past, its being accessed by means of an existential analytic, rather than a "what", like some material object in space'."[149]

Nevertheless, Owen[150] has adopted the new term of 'neuro-agency' that illustrates the "standard term 'agency' in order to acknowledge the role of neurons in human free will".[151]

[141] Keith Farnsworth, 'Can a Robot Have Free Will?' pg1, (2017) Entropy MDPI < https://pure.qub.ac.uk/portal/files/130713217/entropy_19_00237_v3.pdf > accessed on 1 November 2017.
[142] Keith Farnsworth, 'Can a Robot Have Free Will?' (2017) Entropy MDPI < https://pure.qub.ac.uk/portal/files/130713217/entropy_19_00237_v3.pdf > accessed on 1 November 2017.
[143] Keith Farnsworth, 'Can a Robot Have Free Will?' (2017) Entropy MDPI < https://pure.qub.ac.uk/portal/files/130713217/entropy_19_00237_v3.pdf > accessed on 1 November 2017.
[144] J. Basl, 'Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines,' (2014) Journal of Philosophy and Technology, 27(1), 79-96. 2014 < https://philpapers.org/archive/BASMAM.pdf > accessed on 10th December 2018.
[145] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).
[146] Martin Heidegger was a German philosopher (1889–1976).
[147] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017), pg96.
[148] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017), pg96.
[149] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017), pg96.
[150] Tim Owen, 'Cyber-Violence: Towards a Predictive Model, Drawing upon Genetics, Psychology and Neuroscience' (2016) Pg2 < *clok.uclan.ac.uk/16131/1/40256-50104-1-PB-1.pdf*> accessed on 6 November 2016
[151] Tim Owen, 'Cyber-Violence: Towards a Predictive Model, Drawing upon Genetics, Psychology and Neuroscience' (2016) < *clok.uclan.ac.uk/16131/1/40256-50104-1-PB-1.pdf*> accessed on 6 November 2016

Owen[152] defines agency as being affected by morality, reason and inherited constitutional variables, which humans find harder to associate with AMs, although Anderson and Anderson[153] see this as achievable for AMs through the coding of “general ethical principles,” periodic updates, and training to avoid ethical dilemmas.

2.4.3 Autonomy and AM/MAMs

The DoD[154] see the term ‘autonomy’ implying “bounded independent thought and action.”[155] The DoD draw from Herbert in saying:

> “As a fundamental principle, Simon’s Law of Bounded Rationality[156] states that the actions of a program or robot are bounded by the information it has, the amount of time available for computation and the limitations of its algorithms—thus, the independence of a [AM] is fixed by the designers.”[157]

In contrast to the DoD, Eidenmuller[158] believes that ‘smart robots’ are not only able to tell purposive actions, but they can, “exhibit ‘moral agency’: they seem to understand the consequences of their behaviour, and they have a choice of actions.”[159] Eidenmuller[160] acknowledges that the legal personality of robots presents fundamental philosophical problems, which he terms “deep normative structure,” and our acceptance of robot legal

[152] Tim Owen, ‘Cyber-Violence: Towards a Predictive Model, Drawing upon Genetics, Psychology and Neuroscience’ (2016) < *clok.uclan.ac.uk/16131/1/40256-50104-1-PB-1.pdf*> accessed on 6 November 2016

[153] Susan Anderson and Michael Anderson, ‘The Consequences for Human Beings of Creating Ethical Robots,’ (2007) aaai.org < https://www.aaai.org/Papers/Workshops/2007/WS-07-07/WS07-07-001.pdf> accessed on 20 October 2017

[154] US Department of Defense.

[155] Department of Defense Task Force Report: The Role of Autonomy in DoD System (2012) Department of Defense < https://fas.org/irp/agency/dod/dsb/autonomy.pdf > accessed on 5 November 2017.

[156] Herbert Simon, *The Sciences of the Artificial* (3rd edn, MIT Press 1996).

[157] Department of Defense Task Force Report: The Role of Autonomy in DoD System (2012) Department of Defense < https://fas.org/irp/agency/dod/dsb/autonomy.pdf > accessed on 5 November 2017.

[158] Horst Eidenmüller, ‘Robots’ Legal Responsibility’ (University of Oxford, 08 Mar 2017) < https://www.law.ox.ac.uk/business-law-blog/blog/2017/03/robots%E2%80%99-legal-personality > accessed on 1 November 2017.

[159] Horst Eidenmüller, ‘Robots’ Legal Responsibility’ (University of Oxford, 08 Mar 2017) < https://www.law.ox.ac.uk/business-law-blog/blog/2017/03/robots%E2%80%99-legal-personality > accessed on 1 November 2017.

[160] Horst Eidenmüller, ‘Robots’ Legal Responsibility’ (University of Oxford, 08 Mar 2017) < https://www.law.ox.ac.uk/business-law-blog/blog/2017/03/robots%E2%80%99-legal-personality > accessed on 1 November 2017.

personality is, “based on a utilitarian conception of ‘the good’ or whether it rather is based on a humanitarian/Kantian vision according to which not everything that is utility-maximizing is necessarily the better policy.”[161] However, Anderson and Anderson[162] see the development of machine ethics as helping us to understand the meaning of behaving ethically and will therefore advance research into ethics theory.

2.4.4 Ethics

Basl[163] and Hooker[164] both explore machine ethics and our duties and responsibilities, including moral and ethical responsibilities, to an AM, e.g., respecting an AMs autonomy, which is in line with an AM having consciousness. Hooker gives the example of murder and it being contrary to autonomy, consequently posing the idea that we may have a duty to repair AMs rather than destroy them, even if we no longer have a need for them. Hooker[165] argues that AMs could be ethical machines as, “they will be beholden first and foremost to ethics”[166] and wants us to rid ourselves of the idea they will be out the control of humans, an idea supported by Anderson and Anderson.[167] Hooker[168] further believes we will be able to preprogramme much of their behaviour and even implant any cultures and/or personalities

[161] Horst Eidenmüller, ‘Robots’ Legal Responsibility’ (University of Oxford, 08 Mar 2017) < https://www.law.ox.ac.uk/business-law-blog/blog/2017/03/robots%E2%80%99-legal-personality > accessed on 1 November 2017.

[162] Susan Anderson and Michael Anderson, ‘The Consequences for Human Beings of Creating Ethical Robots,’ (2007) aaai.org < https://www.aaai.org/Papers/Workshops/2007/WS-07-07/WS07-07-001.pdf> accessed on 20 October 2017.

[163] J. Basl, ‘Machines as moral patients we shouldn’t care about (yet): The interests and welfare of current machines,’ (2014) Journal of Philosophy and Technology, 27(1), 79-96. 2014 < https://philpapers.org/archive/BASMAM.pdf > accessed on 10th December 2018.

[164] J Hooker, ‘Autonomous Machines Are the Best Kind, Because They Are Ethical’, (2016), Carnegie Mellon University < http://public.tepper.cmu.edu/jnh/agencyPost2.pdf > Accessed on 14 December 2017.

[165] J Hooker, ‘Autonomous Machines Are the Best Kind, Because They Are Ethical’, (2016), Carnegie Mellon University < http://public.tepper.cmu.edu/jnh/agencyPost2.pdf > Accessed on 14 December 2017.

[166] J Hooker, ‘Autonomous Machines Are the Best Kind, Because They Are Ethical’, pg11, (2016), Carnegie Mellon University < http://public.tepper.cmu.edu/jnh/agencyPost2.pdf > Accessed on 14 December 2017.

[167] Susan Anderson and Michael Anderson, ‘The Consequences for Human Beings of Creating Ethical Robots,’ (2007) aaai.org < https://www.aaai.org/Papers/Workshops/2007/WS-07-07/WS07-07-001.pdf> accessed on 20 October 2017.

[168] J Hooker, ‘Autonomous Machines Are the Best Kind, Because They Are Ethical’, (2016), Carnegie Mellon University < http://public.tepper.cmu.edu/jnh/agencyPost2.pdf > Accessed on 14 December 2017.

we wish. Whilst it may be argued his viewpoint sound naive in the face of DL, his paper is dated 2016, so he makes his conclusions in the full knowledge of DL.

Norbert Wiener[169] says, AMs are ‘helpless’, in that they are not in any way an agent and lack the capacity to be ‘moved by reasons’ as Kant states.[170] Thus, for Wiener AMs are not viewed as constrained or restricted agents, although this view will not hold well in a future of AM consciousness. Indeed, the ability to oppose the principle of not regarding a perceived person as a real person is dangerous and obnoxious, smacking of racism, transphobia or speciesism.

Pagallo[171], Duffy and Hopkins[172] and Sandberg[173] draw a parallel between AMs and animals, with Pagallo[174] even suggesting a duty to nurture AMs, very much like a parent does for a child. Pagallo[175] raises some very interesting legal questions such as could we, the purchaser, become liable to others due to not giving the machine the correct stimuli or teaching it ‘bad’[176] behaviour? He strongly hints that this could be a future possibility, reinforced by Basl[177], but contested by Owen and Owen[178] who see the manufacturer as always responsible. Teaching our values and IHL principles to AM/MAMs, helps ensure they understand and

[169] Norbert Wiener (1894 – 1964) was an American mathematician and philosopher and credited as being one of the first to theorize about machine intelligence.

[170] Garrath Williams, ‘Kant’s Account of Reason’, The Stanford Encyclopedia of Philosophy (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), <https://plato.stanford.edu/archives/fall2023/entries/kant-reason/> accessed 10 September 2023.

[171] Ugo Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts* (2013 edn, Springer).

[172] Sophia Duffy and Jamie Hopkins, ‘Sit, Stay, Drive: The Future Of Autonomous Car Liability” (2014) < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2379697 > accessed on 10 December 2018.

[173] Anders Sandberg, ‘Law-abiding robots?’, (2016) University of Oxford < https://www.oxfordmartin.ox.ac.uk/opinion/view/340 > accessed on 24 October 2017.

[174] Ugo Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts* (2013 edn, Springer).

[175] Ugo Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts* (2013 edn, Springer).

[176] This is very subjective.

[177] J. Basl, ‘Machines as moral patients we shouldn’t care about (yet): The interests and welfare of current machines,’ (2014) Journal of Philosophy and Technology, 27(1), 79-96. 2014 < https://philpapers.org/archive/BASMAM.pdf > accessed on 10th December 2018.

[178] Tim Owen, Crime, Genes, Neuroscience and Cyberspace (Palgrave Macmillan 2017).

comply, which address the VAP. Yet the duty to nurture also raises what the author sees as a fundamental gaps and terms the true value alignment problem (TVAP), which is broadening and aligning our values to include, and for the benefit of, AM/MAMs.

When it comes to AMs making decisions for themselves and/or on behalf of us, Bonnefon et al[179] conducted a survey into utilitarian autonomous vehicles (AVs) and the views of saving more lives even if it means self-sacrificing. They found that the respondents were supportive of utilitarian AVs (utilitarian AMs is an idea supported by Hooker)[180] and even supportive of self-sacrificing when faced with killing others, however respondents were opposed to making self-sacrifices legal enforceable, which Bonnefon et al[181] know would provoke significant ethical arguments. Bonnefon et al[182] believe polling the public will, "inform the construction and regulation of moral algorithms for AVs, not dictate them,"[183] which can be transposed to all other AMs with the possible exception of MAMs. Indeed, Human Rights Watch (HRW) fight for "meaningful human control",[184] which they see as a vital condition to ensure human control of weapons and consequently removing the majority of issues related with fully

[179] Jean-Francois Bonnefon, Azim Shariff and Iyad Rahwan, 'Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?' (2015) Researchgate.net < https://www.researchgate.net/publication/282843902_Autonomous_Vehicles_Need_Experimental_Ethics_Are_We_Ready_for_Utilitarian_Cars > accessed on 24 October 2017.

[180] J Hooker, 'Autonomous Machines Are the Best Kind, Because They Are Ethical', (2016), Carnegie Mellon University < http://public.tepper.cmu.edu/jnh/agencyPost2.pdf > Accessed on 14 December 2017.

[181] Jean-Francois Bonnefon, Azim Shariff and Iyad Rahwan, 'Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?' (2015) Researchgate.net < https://www.researchgate.net/publication/282843902_Autonomous_Vehicles_Need_Experimental_Ethics_Are_We_Ready_for_Utilitarian_Cars > accessed on 24 October 2017.

[182] Jean-Francois Bonnefon, Azim Shariff and Iyad Rahwan, 'Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?' (2015) Researchgate.net < https://www.researchgate.net/publication/282843902_Autonomous_Vehicles_Need_Experimental_Ethics_Are_We_Ready_for_Utilitarian_Cars > accessed on 24 October 2017.

[183] Jean-Francois Bonnefon, Azim Shariff and Iyad Rahwan, 'Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?' (2015) Researchgate.net < https://www.researchgate.net/publication/282843902_Autonomous_Vehicles_Need_Experimental_Ethics_Are_We_Ready_for_Utilitarian_Cars > accessed on 24 October 2017.

[184] Human Rights Watch, 'Killer Robots and the Concept of Meaningful Human Control' (2016) <https://www.hrw.org/news/2016/04/11/killer-robots-and-concept-meaningful-human-control > accessed on 6 November 2016.

autonomous weapons, which includes the value alignment problem (VAP)[185]. HRW argue that, "such a requirement would protect the dignity of human life, facilitate compliance with international humanitarian and human rights law, and promote accountability for unlawful acts."[186] Disagreeing, Hooker[187] does not see AMs as taking over or violating our ethical rules, as this would go against the principle of autonomy. Nevertheless, HRW are resolute in their belief that a human should always maintain control over weapons of war, arguing humans are inclined, "to feel the emotional weight and psychological burden of choosing to take away the life of other human beings…Empathy can act as a check on killing, but only if humans have control over whom to target and when to fire."[188]

In 2012 and under the Obama administration, the Pentagon issued the United States Department of Defense (DoD) Policy Directive 3000.09 for the development and use of weapon systems that were both semi-autonomous and fully autonomous. Most notably, the policy states "appropriate levels of human judgment"[189] are necessary for deciding on the force required, although no definition offered is offered. Nevertheless, despite human judgement, there are examples of things going wrong, for example the 1988 US incident where a Navy air defense system shot down an Iranian airliner by mistake. Human error was a contributing factor, however a significant part of the problem appeared to be an over reliance on the automated systems. False trust of the autonomous weapon's judgment and

[185] Brian Christian, *The Alignment Problem: How Can Artificial Intelligence Learn Human Values?* (September 2021, Atlantic Books)

[186] Human Rights Watch, 'Killer Robots and the Concept of Meaningful Human Control' (2016) <https://www.hrw.org/news/2016/04/11/killer-robots-and-concept-meaningful-human-control > accessed on 6 November 2016.

[187] J Hooker, 'Autonomous Machines Are the Best Kind, Because They Are Ethical', (2016), Carnegie Mellon University < http://public.tepper.cmu.edu/jnh/agencyPost2.pdf > Accessed on 14 December 2017.

[188] Human Rights Watch, 'Killer Robots and the Concept of Meaningful Human Control' (2016) <https://www.hrw.org/news/2016/04/11/killer-robots-and-concept-meaningful-human-control > accessed on 6 November 2016.

[189] US Department of Defense Directive NUMBER 3000.09 November 21, 2012 < https://fas.org/irp/doddir/dod/d3000_09.pdf > accessed on 24 October 2017.

capability can lead to the human soldiers over reliance on the weapon that could lead to devastating consequences.

The degree of meaningful human control deployed for non-conscious MAMs is yet to be truly established. Consequently, the conscious MAMs, which are the focus of this research, will remove all control and place the MAM on the same footing as a human soldier. Galliott et al[190] view the 'designers' of MAMs a creating the ethico-legal gap, which they suggest is filled through a process called "Value Sensitive Design",[191] but this still views humans as the originators of MAM actions. No where do they consider MAMs being conscious, deciding their own course of action and the impact this will have. This is where the VAP starts to emerge in relation to IHL. The VAP looks at how AM/MAMs are developed to behave and act in accordance with human norms and values.[192] Nevertheless, MAMs could in fact increase our compliance to the values of IHL, as discussed throughout the research questions.

2.4.5 Value Alignment

Trust is key to our acceptance of AM/MAMs, and their ability to align to our values. There appears to be a distrust of AMs[193], although Holder et al[194] strongly evidence against this. Carlson et al[195] aim to counter any distrust by developing models of the aspects that impact

[190] Jai Galliott, Duncan MacIntosh, and Jens David Ohlin, *Lethal Autonomous Weapons. Re-Examining the Law and Ethics of Robotic Welfare* (2021 Oxford University Press).
[191] Jai Galliott, Duncan MacIntosh, and Jens David Ohlin, *Lethal Autonomous Weapons. Re-Examining the Law and Ethics of Robotic Welfare* (2021 Oxford University Press).
[192] Brian Christian, *The Alignment Problem: How Can Artificial Intelligence Learn Human Values?* (September 2021, Atlantic Books).
[193] Andrew Bolster and Alan Marshall, 'A Multi-Vector Framework for Autonomous Systems' (2014) www.AAAI.org < https://www.researchgate.net/publication/271699529_A_Multi-Vector_Trust_Framework_for_Autonomous_Systems > accessed on 10 December 2017.
[194] Chris Holder, Vikram Khurana, Faye Harrison, and Louisa Jacobs, 'Robotics and law: Key legal and regulatory implications of the robotics age (Part I of II)' [2016] Computer Law & Security Review, Volume 32, Issue 3.
[195] Michelle Carlson, Munjal Desai,Jill Drury, Hyangshim Kwak and Holly Yanco, 'Identifying Factors that Influence Trust in Automated Cars and Medical Diagnosis Systems' (2013) AAAI Spring Symposium <

our trust in AMs, which they hope designers will use to develop AMs that we can trust and align. They admit that they still need to undertake more research to develop the models. Lockheed Martin[196] support the DoD's ambition to develop autonomous military machines to protect soldiers in the battlefield and are just as quick to steer the conversation away from the controversial side and 'accidental wars'. Further, MAMs, which will operate in situations of conflict, must adhere to ethical and legal norms, that are borne out of the atrocities and evils of days gone by. War should always be considered as an absolute last resort. It is never pleasant and kills and injures many soldiers and civilians. It devastates a country's infrastructure, its cultures, and communities. Poverty worsens and valuable environmental resources are ruined. These consequences are wrapped up and feed into just war theory, which is discussed shortly.

When considering conflict, which State is right or wrong is sometimes difficult to determine, yet the State's beliefs validate their war over another State as just and reasonable due to jurisdiction issues.[197]

Military techniques and methods can be judged inappropriate and unethical (e.g., waterboarding) and the certain methods are prohibited such as genocide, rape, and torture.

The Hague Convention of 1907 prohibits the use of:

http://robotics.cs.uml.edu/fileadmin/content/publications/2014/CarlsonDesaiDruryKwakYanco-Trust-SSS14.pdf > accessed 24 October 2017.

[196] US defence company who have development many military machines and aircraft, and who have just developed the Autonomous Mobility Applique System (AMAS).

[197] Bertrand Russell, "The Ethics of War" (1915) International Journal of Ethics, vol. 25, no. 2, 1915, pp. 127–42 < http://www.jstor.org/stable/2376578 > accessed 24 May 2019.

- Poison or poisoned weapons
- Killing or wounding treacherously
- Killing or wounding an enemy who has surrendered
- Stating that no mercy will be granted to defeated opponents
- Using arms, projectiles, or material designed to cause unnecessary harm and suffering

Realist ontology[198] accepts that values and norms significantly differ between civilian and military environments. The military setting is a unique environment with its own culture, values and norms. Within the civilian environment, we are expected to view everyone as equal, not harm anyone or any property, embrace difference and be tolerant of behaviours. That being said, there is a strong suspicion around the application of moral judgements and concepts to the management of international affairs.[199] Realists unite on the importance of power and security issues, on the perception that State's need to maximise their self-interest and, paramount, their view that the international stage is simply one of anarchism, which should only be restored if it is in the State interest. Further, realists emphasise that, once a war has begun, a State should take whatever action it can to win.

2.4.6 Just War Theory

Kant[200], who is frequently referred to as the first truly international political philosophers, is often considered as not having a just war theory (JWT) and, furthermore, was a ferocious

[198] Niiniluoto, Ilkka, 'Realism in Ontology' (2003) Critical Scientific Realism, Oxford Academic, 1 Nov. 2003 < https://doi.org/10.1093/0199251614.003.0002 > accessed 4 September 2024.
[199] Brian Orend (2004) Kant's ethics of war and peace, Journal of Military Ethics, 3:2, 161-177.
[200] Immanuel Kant was a German philosopher, born 1724.

critic of classical just war theorists, such as Augustine, Aquinas, and Grotius. Kant[201] views States as creating deliberate choices and policies, rather than being determined by realpolitik. Moreover, Kant,[202] does not accept war as a prevalent certainty of international life. In fact, Kant[203] asserts that States can and should act morally, with us all judging their moral behaviour,[204] which indeed is the with IHL. It could be thought that Kant snubs the ideas of realism, as Kant does not advocate basing foreign policy on a foolish approach of exploiting national interests, rather advocating regarding the needs of justice. Indeed, "Kant is adamant that the purely prudential approach to foreign policy is 'immoral and opportunistic'". [205] A wholly 'prudential' foreign policy is disloyal to, "our most fundamental identity as rational beings responsive to the demands of morality and justice." [206] This feeds into the principles of IHL, specifically necessity, and the stipulation that " state engaged in an armed conflict is permitted to use only that degree and kind of force, not otherwise prohibited by the law of armed conflict, that is required in order to achieve the legitimate purpose of the conflict."[207]

Just war theory and ethics is part of moral theology, and forms part of the ethical consciences of those involved and touched my military policies and conflict, but it is not part of international law.[208] It was first developed by St Thomas Aquinas[209], and has origins in Christian philosophy. It tries to conjoin three complex thoughts, although can be used by anyone with or without faith. The core principles of JWT, which Christianity built upon,

[201] Immanuel Kant was a German philosopher (22 April 1724 – 12 February 1804).
[202] Immanuel Kant was a German philosopher (22 April 1724 – 12 February 1804).
[203] Immanuel Kant was a German philosopher (22 April 1724 – 12 February 1804).
[204] It should be noted that Kant's arguments do contradict themselves, as States are not actors. States cannot form decisions, are not conscious and do not have personhood.
[205] Brian Orend (2004) Kant's ethics of war and peace, Journal of Military Ethics, 3:2, 161-177.
[206] Brian Orend (2004) Kant's ethics of war and peace, Journal of Military Ethics, 3:2, 161-177.
[207] UK Parliament, 'The Law Governing Armed Conflict' (Parliamentary Business, 22 July 2019), < https://publications.parliament.uk/pa/cm201719/cmselect/cmdfence/1224/122405.htm > accessed 4 September 2024.
[208] Joseph C. Sweeney, 'The Just War Ethic in International Law' (2003) 27 Fordham Int'l L.J. 1865 < https://ir.lawnet.fordham.edu/ilj/vol27/iss6/2 > accessed 4 September 2024
[209] Thomas Aquinas (ca. 1225–1274) was a philosopher and theologist.

originated with classical Greek and Roman philosophers, such as Plato and Cicero, then further developed by Christian theologians such as Augustine and Thomas Aquinas. JWT fundamentally believes that:

- Taking a human life is wrong.
- States have a duty to defend their citizens and their justice.
- The protection of innocent human life and the defending of moral values at times requires readiness to use force and violence.

The purpose of JWT is to offer guidance to the 'right' way for states to act when faced with potential conflict. It applies to states only, so not to individuals, however an individual could use the theory to help them consider if it is morally right to take part in a war. It provides a valuable framework for state leaders and political groups to discuss potential wars. It is designed to prevent wars rather than be used to justify them, by highlighting other ways to resolve conflict. However, JWT can mislead a state into judging that because a war is just, it is then in fact good. A just war may be permissible, but it's still a war where lives will be lost.

There are two core parts to JWT:[210]

[210] M Hadji-Janev, and K Hristovski, 'BEYOND THE FOG: AUTONOMOUS WEAPON SYSTEMS IN THE CONTEXT OF THE INTERNATIONAL LAW OF ARMED CONFLICTS' (2017) Jurimetrics Journal of Law, Science and Technology, 57(3), 325+, < https://link.gale.com/apps/doc/A517345660/ITOF?u=griffith&sid=bookmark-ITOF&xid=2288b33d > accessed 6 Sep 2019.
There is a third dimension to JWT, which is jus post bellum, meaning justice after war and encompasses the transition from war to peace.

- Jus ad bellum: The conditions under which a state may go to war and therefore the use of military force is justified; and
- Jus in bello: The conduct of parties in armed conflict (war conducted in an ethical manner).

A war can only be considered a just war and ethical if both of the above core conditions are met; justified and undertaken in the right way. Wars that have been fought for noble causes have subsequently been judged unjust due to the way in which they were fought. As Russell back in 1915 stated, "It is necessary, in regard to any war, to consider, not its paper justification in past agreements, but its real justification in the balance of good which it is to bring to mankind."[211]

It should be noted that, according to Cook & Syse[212], nobody is the exclusive authority on military ethics, be they a military lawyer or a military chaplain.[213] Military ethics offers a "conceptual framework by means of which one can assess the value of the various military ethics activities,"[214] with critical assessment of the Laws of Armed Conflict (LOAC) as, "a fundamental component of military ethics, understood as professional ethics."[215] Therefore, research aim 3 will discuss the emerging ethical questions and the ethical challenges MAM

[211] Bertrand Russell, "The Ethics of War" (1915) International Journal of Ethics, vol. 25, no. 2, 1915, pp. 127–42 < http://www.jstor.org/stable/2376578 > accessed 24 May 2019

[212] Martin L. Cook & Henrik Syse, "What Should We Mean by 'Military Ethics'?" [2010] Journal of Military Ethics, 9:2, 119-122.

[213] Martin L. Cook & Henrik Syse, "What Should We Mean by 'Military Ethics'?" [2010] Journal of Military Ethics, 9:2, 119-122.

[214] Martin L. Cook & Henrik Syse, "What Should We Mean by 'Military Ethics'?" [2010] Journal of Military Ethics, 9:2, 119-122.

[215] Martin L. Cook & Henrik Syse, "What Should We Mean by 'Military Ethics'?" [2010] Journal of Military Ethics, 9:2, 119-122.

decision making could present IHL, along if our values will need reviewing and expanding to address the author’s identified TVAP.

2.5 Research Aim 4: To understand whether MAMs truly align to the principles of IHL

2.5.1 Alignment Challenges

The advent of “thinking machines”[216] cited by Bostrom & Yudkowsky[217] and their related ethical problems, yield legal dilemmas and challenges for the law and alignment of IHL principles. Savirimuthu affirms that “the role of law in these debates is particularly conspicuous by its perceived inability to keep pace with emerging technologies and concerns.”[218] Further, Savirimuth emphasises that “our understanding of rules and decision making has long been based on normative frameworks and standards involving human–human interactions in society.”[219] He identifies that questions need to answered about how we transpose or extend existing frameworks, values and standards to “non-biological entities” such as AMs, especially when these rules can be vague, open to interpretation and situation specific, which again feeds into the VAP. However, the literature does not discuss how we extend our values to AMs, which the author has termed here the ‘true value alignment problem’ (TVAP). The TVAP becomes of increased importance when exploring MAMs and the principles of IHL

[216] Nick Bostrom and Eliezer Yudkowsky, ‘The Ethics of Artificial Intelligence’ [2011] Cambridge University Press.

[217] Nick Bostrom and Eliezer Yudkowsky, ‘The Ethics of Artificial Intelligence’ [2011] Cambridge University Press.

[218] Savirimuthu, Joseph, “Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence” Patrick Lin, Keith Abney and Ryan Jenkins (eds), International Journal of Law and Information Technology, Volume 26, Issue 4, Winter 2018, Pages 337–346, <https://doi.org/10.1093/ijlit/eay011> accessed 9 January 2023.

[219] Savirimuthu, Joseph, “Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence” Patrick Lin, Keith Abney and Ryan Jenkins (eds), International Journal of Law and Information Technology, Volume 26, Issue 4, Winter 2018, Pages 337–346, <https://doi.org/10.1093/ijlit/eay011> accessed 9 January 2023.

To date, very little regard has been given to conscious AMs and what has been contemplated has been framed generally as thought experiments and not deemed as possible future reality. This limited regard endures when deliberating the rights and duties we may all owe AMs and the unique circumstances of MAMs. A question to debate and answer before we process with AM technology and create artificially conscious AMs, is 'do we actually need artificial conscious agents?' We have an excess of natural conscious agents that can undertake tasks, and which have sometimes been marginalised, e.g., people with disabilities. Conversely, what we do need is intelligent tools to facilitate us. We know that children are not able to enter contracts except in specific cases of necessity, and people without capacity are under the legal care and responsibility of guardians. Thus, research aim 4 will discuss the TVAP and the need for humans to review the principles under IHL, with extending them to protect MAMs in mind.

The literature for each research aim will now be discussed in more detail with the each chapter.

3. RA1: To explore and understand the legal landscape of AM and IHL.

3.1 Introduction

> "For the survivor who chooses to testify, it is clear: his duty is to bear witness for the dead and for the living. He has no right to deprive future generations of a past that belongs to our collective memory. To forget would be not only dangerous but offensive; to forget the dead would be akin to killing them a second time." Elie Wiesel, Holocaust survivor.[220]

Back in the 4th Century, Flavius Vegetius Renatus wrote the Roman warfare and military principles[221], where the Latin phase 'si vis pacem, para bellum', which translates to 'if you want peace, prepare for war', was founded. Today, with the development of AMs and MAMs, this holds more prominence and sounds as a warning for us to proactively plan ahead for their integration into our lives and deployment in warfare.

The law tends to derive from societal beliefs, norms, and values[222], which have an ethical root and encapsulates the value alignment problem (VAP). Therefore, this research aim (RA) links with the discussion in RA3 regarding value, beliefs, rights and duties, showing how the law has evolved to address trends and movements. Moreover, with the dawn of AMs, the law is facing an ever-growing challenge to keep pace. In exploring the challenges the law will face with AMs, English criminal law is initially explored, as the core elements of criminal liability

[220] Quote from Holocaust survivor Elie Wiesel describing his experiences in the Auschwitz and Buchenwald concentration camps: Elie Wiesel, 'Night,' [1972] Penguin Books.

[221] Flavius Vegetius Renatus, De Re Militari, 1473.

[222] P. Sales, 'CONSTITUTIONAL VALUES IN THE COMMON LAW OF OBLIGATIONS' (2024) The Cambridge Law Journal, 83(1) < https://www.cambridge.org/core/journals/cambridge-law-journal/article/constitutional-values-in-the-common-law-of-obligations/10695D32CEDAA3E2EEC391C212AF3925 > accessed 4 September 2024.

(mens rea and actus reus), which have a high evidential threshold, and must be proved for criminal liability. Consequently, with AMs, it is argued in this thesis that they will meet the criminal liability requirements. Further, there are situations where military personal, and thus MAMs, could be tried for criminal liability, due to acting outside of International Humanitarian Law (IHL), therefore, just war theory (JWT) (introduced in chapter 2) and IHL are examined.

3.2 Legal Framework

IHL[223], also known as the Law of Armed Conflict (LOAC), is the legal framework applicable to situations of armed conflict and occupation, of which all UN member states abide by. The MoD see the LOAC as governing not only the conflict, but the restoration effort, so all-encompassing and wide ranging. Those countries that have signed a convention governing warfare, understand that their military will be punished if the break any of the convention rules.

When discussing IHL, the International Committee of the Red Cross (ICRC) stress:[224]

> "IHL applies to the belligerent parties irrespective of the reasons for the conflict or the justness of the causes for which they are fighting…IHL is intended to protect victims of armed conflicts regardless of party affiliation. That is why jus in bello must remain independent of jus ad bellum… IHL is purely humanitarian, seeking to limit the suffering caused."[225]

[223] Also known as the 'laws of war' or the 'law of armed conflict' (LOAC).

[224] The conventions and laws governing warfare are discussed in detail in chapter 4, Law.

[225] ICRC, 'What are jus ad bellum and jus in bello?, (ICRC, 22 January 2015) <https://www.icrc.org/en/document/what-are-jus-ad-bellum-and-jus-bello-0 > accessed 2 October 2019.

The just war practice has often been seen as evolving between two culturally similar enemies. Thus, when values are aligned between two fighting people, it is frequently found that they implicitly or explicitly come to an understanding of the limits of their warfare. Yet when enemy States clash due to differing values, religious beliefs, status, race, or language, it can feed a view that that each other is "less than human", and as such war conventions are seldom applied. MAMs may alter this as they communicate in a universal language (computer code), and which is discussed further in chapter 4. MAMs may not be attached to values the same way that humans are, nevertheless, they are still an asset of a specific State, so will be indoctrinated with the values and cultures of that State. Consequently, MAMs are developing and learning in the context of what they are exposed to by their home State, and is the responsibility of such, especially when things go wrong, as in the case of Marine A[226].
Marine A is a case (discussed later in this chapter), resulting in criminal liability due to a breach of IHL. Therefore, as a result of the possibility of criminal action for breaches of IHL, criminal liability is explored as it forms part of the legal landscape.

Whilst we hope that AMs will enhance our day-to-day life, for example, monitoring our health and prompting us to take action, there are environments where the use of AMs will be high risk and contentious, such as the military environment. MAMs, which will operate in situations of conflict, and must adhere to the ethico-legal norms, that are borne out of the atrocities and evils of days gone by. We must learn from these and ensure we do everything

[226] R v Sergeant Alexander Wayne Blackman ("Marine A"), Case Reference: 2012CM00442

in our power to teach and control MAMs to behave like a human soldier, if not better; We must learn from history, so we do not sleepwalk into the same or new atrocities.

3.2.1 Exploring the Legal Landscape

3.2.1.1 Civil Law

Part of setting the context, is appreciating what the current legal framework is for determining liability when someone is harmed, or even killed, due to a product. Protection is afforded through Acts and case law derived from criminal and civil areas of law. The Consumer Protection Act 1987 (CPA 1987) is central to product liability, with Parts II and IV covering criminal liability. CPA 1987 holds manufacturers accountable for safety and quality, and imposing strict liability, therefore negligence does not need to be proven if a product is defective. Tortious liability common law claims are founded upon Donoghue v Stevenson[227] and further covers third parties[228]. Subsequent legislation, including the Consumer Protection Act 1987 and the General Product Safety Regulations 2005, mandate that products must meet safety standards. Nevertheless, a statutory defence exists for those who can prove they took all reasonable steps to prevent harm. Additionally, the General Product Safety Regulations 2005/1803 is non-sector-specific catch-all for criminal liability and applies if there are not any specific EU rules controlling the safety of a product.

As AI and AMs evolve, it is the concern here that current product liability provisions are unlikely to remain suitable, and may even be deemed obsolete. Traditional concepts of intent (mens rea) and action (actus reus) will be challenged, thus highlighting potential legal challenges in attributing responsibility and culpability.

[227] Donoghue v Stevenson [1932] AC 562.
[228] Lambert v Lewis [1982] AC 225.

Smart and Richardson[229] are very bold in proclaiming;

> "Robots, even sophisticated ones, are just deterministic machines. They will be no more than machines for the foreseeable future, and we should design our legislation accordingly. Falling into the trap of anthropomorphism, or even ascribing to a machine anima that is simply not there, will lead to contradictory situations, bad legislation, and bad court decisions. It will be bad for society and bad for the nascent robotics industry."[230]

Perhaps their lack of foresight or vision could be due to publishing their paper in 2014, as technology has advanced significantly in the subsequent years. This view is not shared by Solaiman, who appears to be a forward thinker by recognising the legal holes that will appear with the advancement of AMs.[231] However, Solaiman deems AMs could not be compared with animals, "in terms, for instance, of so-called autonomy, self-awareness, or self-determination, though the latter may be more autonomous compared to the former."[232] Despite an AM being human-made and an animal being a living biological animal, Solaiman and Leenes and Lucivero[233] regarded AMs as property and consequently 'objects' of law, which is aligned to product legislation, rather than being 'subjects'. Thus, they view a human is always legally responsible for the AM's actions and therefore responsible for guaranteeing that they operate within the confines of the law. Subsequently, the human is being extrinsically controlled[234], not the AM, yet it is the view here that this will be impracticable with AM/MAMs.

[229] Neil M Richards and William D Smart, 'How Should the Law Think About Robots?' (10 May 2013) < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2263363> accessed on 17th November 2017.

[230] Neil M Richards and William D Smart, 'How Should the Law Think About Robots?' (10 May 2013) < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2263363> accessed on 17th November 2017.

[231] S.M Solaiman, 'Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy' (2017) Artif Intell Law 25, 155–179 (2017) <https://link.springer.com/content/pdf/10.1007/s10506-016-9192-3.pdf > accessed on 17th October 2018.

[232] S.M Solaiman, 'Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy' (2017) Artif Intell Law 25, 155–179 (2017) <https://link.springer.com/content/pdf/10.1007/s10506-016-9192-3.pdf > accessed on 17th October 2018.

[233] R Leenes and F Lucivero, 'Laws on robots, laws by robots, laws in robots: regulating robot behaviour by design' [2014] Law Innov Technol 6(2).

[234] A.B.A Majeed, 'Roboethics - Making Sense of Ethical Conundrums' (2017) Procedia Computer Science, Volume 105, 2017 < https://doi.org/10.1016/j.procs.2017.01.227 > accessed 19 March 2019.

Where a non-autonomous machine malfunctions causing harm, then attributing product and/or criminal liability to a human or the manufacturer, is established through the wide body of law as appropriate. When comparing AMs with pets, there is an argument for owner's liability, as confirmed by Weng et al[235] and Leenes and Lucivero[236], whilst Bertolini[237] is supportive of manufacturers' liability covered by product liability. Liability is viewed as an incentive as opposed to a deterrent, in that it encourages fear in potential offenders, as a result creating compliance with the law and to the prevention of harm. Solaiman, whom is against assigning legal personhood to AMs[238], warns that, "granting legal personality to robots may not be a panacea; rather it may turn out to be Pandora's box, if we transform the machines to our masters."[239] It is worth reiterating that assigning legal personality to AMs risks absolving humans of liability and, in doing so, reducing the effectiveness of deterrence. Bryson offers a sobering thought that, "we are obliged not to the robots, but to our society. We are obliged to educate consumers and producers alike to their real obligations with respect to robotics."[240] However, this is not a reason to ignore the development of AMs and fails to address the liability risks. In reality, it is our obligation to society to address to safeguard legal process and redress.

To date, these product liability frameworks have served us well, however, when examining these laws, and at the time of writing, none specifically cover unconscious AMs, let alone

[235] YH Weng, CH Chen and CT Sun CT, 'Toward the human–robot co-existence society: on safety intelligence for next generation robots' [2009] Int J Social Robot 1:267–282.
[236] R Leenes and F Lucivero, 'Laws on robots, laws by robots, laws in robots: regulating robot behaviour by design' [2014] Law Innov Technol 6(2).
[237] A Bertolini, 'Robots as products: the case for a realistic analysis of robotic applications and liability rules' [2013] Law Innov Technol 5(2):214–247.
[238] Defined and discussed in Chapter 5.
[239] S.M Solaiman, 'Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy' (2017) Artif Intell Law 25, 155–179 (2017) <https://link.springer.com/content/pdf/10.1007/s10506-016-9192-3.pdf > accessed on 17th October 2018.
[240] JJ Bryson, 'Robots should be slaves. In: Wilks Y (ed) Close engagements with artificial companions: key social, psychological, ethical and design issue' [2010] John Benjamins Publishing Company, Amsterdam.

conscious AMs, which is a significant concern and gap. AM technology is rapidly advancing with support from key industry and influencing bodies, (e.g., governments, Google), so it is now for the law to look at futureproofing itself to cover the new challenges that will arise.

The research in chapter 2 highlights that AM legal personhood is very much a matter we need to resolve, as this impacts the relationship with the law. There currently exists, "a tension as to whether AI [of which AM development derives] should be treated as an object, a subject, a thing or a person…reformulating our relationship with AI in a more radical fashion."[241] Indeed the EU draft plan[242] explores and promotes the options of turning AMs into electronic persons. Consequently, it is important to recall Asaro's[243] view of AMs, saying, "autonomous artificial agents can act in the world independently of their designers and operators"[244] and he stresses they will act upon their own decisions, with unpredictable and unintended actions and effects. Savirimuthu warns that whilst the law tries to get a grip of the technology, then translate this into effective legislation, individuals "will be treated as laboratory experiments when sensors do not function as expected, or if a robot is hacked or corrupted by malware."[245] This is likely to be considered inappropriate and to be met with legal claims.

[241] Jacob Turner, 'Robot Rules. Regulating Artificial Intelligence', 2019, Palgrave Macmillan.

[242] European Parliament, '*DRAFT REPORT with recommendations to the Commission on Civil Law Rules on Robotics'* (2015/2103(INL)).

[243] Peter Asaro, 'The Liability Problem for Autonomous Artificial Agents' [2016] Association for the Advancement of Artificial Intelligence.

[244] Peter Asaro, 'The Liability Problem for Autonomous Artificial Agents' [2016] Association for the Advancement of Artificial Intelligence.

[245] Savirimuthu, Joseph, "Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence" Patrick Lin, Keith Abney and Ryan Jenkins (eds), International Journal of Law and Information Technology, Volume 26, Issue 4, Winter 2018, Pages 337–346, <https://doi.org/10.1093/ijlit/eay011> accessed 9 January 2023.

3.2.1.2 Criminal Law

As stressed, manufacturers of AMs are not so keen to be held criminally liable for a machine that self-learns based on stimuli and its environment, all of which is outside of their control. Thus, they look to limit or remove their responsibility wherever possible.

The question of criminal liability is the fundamental question that must be answered in criminal law and requires a higher degree of fault mainly due to the severity of sanctions that can be imposed on a human or corporation. Indeed, should an AM be able to meet key legal elements of a crime, would it be just to hold a manufacturer or user liable? Criminal law is explored here to understand the challenges we may face in ascribing liability to a human who is not deciding or directing the AMs actions. Again, criminal law has the highest evidential bar and can be applied to military personal acting outside of IHL.

Within English law, the age of criminal responsibility is 10 years old.[246] Much has been debated over the nature versus nurture explanations for criminal behaviour, however, the nature-nurture debate has been professed as superfluous, naive and unhelpful by scientists and social scientists, such as Craddock[247] and Traynor[248]. With AMs, the debate changes focus, as AMs will be designed and deployed with a purpose (nature), for example, MAM, whereas babies are not bred to be soldiers. The nurturing or environment an AM 'grows' and develops in will have a significant impact on their behaviour and values, very much like a it has on a human, however, if an AM will be afforded 10 years of development and 'maturing' time

[246] CPS, 'Youth Crime' (CPS, 2022) < https://www.cps.gov.uk/crime-info/youth-crime > accessed 19 March 2019.

[247] N Craddock, 'Horses for courses: the need for pragmatism and realism as well as balance and caution. A commentary on Angel' [2011] Social Science and Medicine 73.

[248] B J Traynor, and AB Singleton, 'Nature versus nurture: death of a dogma, and the road ahead' [2010] Neuron 68.

before being capable of being criminally liable, is argued here to be unacceptable and further, it is the view here that liability will be assigned to them from the moment they are switched on. Environmental rationalisations of behaviour can be portrayed as deterministic and have a negative impact on humans, especially those from disadvantaged backgrounds. It limits social mobility, life expectations and creates class divide. Whilst this may not be the quite the same for AMs, there will be a difference to their behaviour, values and even tolerance dependent upon the environment they are exposed to and the quality of data they have access to, which could relate to the differing levels resources and support within our current social structure.

As mentioned, for there to be criminal liability, the two elements of actus reus and mens rea must be met. If either one of these elements are missing, then no criminal responsibility can be enforced. Further, under English law, there is a presumption of innocence until proven guilty. To enforce criminal responsibility, it must be proven beyond reasonable doubt that the above two elements existed. Thus, if proved that a person committed a criminal act knowingly or with criminal intention, that person is regarded criminally liable for the offence. This basic premise is explored with regards to AMs and their ability to be criminally liable.

3.2.1.3 Actus Reus

Actus reus (AR)[249] is the physical element that must be proved for criminal liability to be imposed. This includes both actions and, in some cases, omissions[250], which are failures to

[249] LexisNexis, 'Criminal act or omission' (LexisNexis, 2024) < https://www.lexisnexis.co.uk/legal/guidance/criminal-act-or-omission > accessed 19 March 2019.

[250] LexisNexis, 'Criminal act or omission' (LexisNexis, 2024) < https://www.lexisnexis.co.uk/legal/guidance/criminal-act-or-omission > accessed 19 March 2019.

act. In general, the law does not impose a legal duty to act, though there may be a moral expectation. Legal liability for omission requires two conditions: (1) the crime can be committed by failing to act, and (2) the defendant has a duty to act, which could arise from contractual obligations[251] or certain relationships[252].

With regards to AMs, those in the military medical arena may well have a duty arising out of the care for another and therefore liable[253] Further, a duty may arise by the making of a dangerous situation[254], for example, a duty to protect a vulnerable person or administer medication. This duty and the resulting consequences, become heightened in the military context, which is explored later. To hold a defendant liable, it must be shown that their actions or omissions caused the consequences in question, both in terms of factual causation[255] and legal causation[256].

3.2.1.3.1 MAM Impact on Causation

Today, a human is the controller of all the acts of unconscious AMs and therefore liable. Yet for the conscious AMs, which are the focus in this thesis, both the legal and factual causations elements could be applied fairly and with little difficulty. The fact that there will be an evidence trail of actions taken due to the nature of the technology, means that it could be easier to assign causation. Indeed, it is the view here that identifying the causation elements

[251] Pittwood (1902) 19 TLR 37.

[252] Gibbins and Proctor (1918) 13 Cr App R 134.

[253] Nicholls (1874) 13 Cox CC 75.

[254] Millar [1983] 2 AC 161.

[255] From the case of White [1910] 2 KB 124. and uses the 'but for' test.

[256] Focuses on the fairness of the outcome (e.g., death) being attributed to the defendant, as in the case of Dalloway (1847) 2 Cox CC 273.

could be made easier to determine due to the following characteristics, which will be of significant importance for MAMs, their decision making, and in ensuring compliance with IHL:

- Evidence Collection and Processing: MAMs will collect and analyse immense amounts of data with precision and speed. For example, a MAM could gather high-quality video footage, sensor data, and environmental information, which could aid in establishing whether a particular action by another solider/State was the factual cause of an event. The same data can be used to understand the decision making, and resulting action, of a MAM.

  Through the recording of real-time detailed information, it is the view here that MAMs offer more reliable, impartial evidence, reducing reliance on subjective witness testimonies, which are often limited by human perception and memory.
- Complex Causal Analysis Through MAM Reasoning: MAMs are thought here as having advanced reasoning abilities, which allow the to model complex causal relationships, that would be challenging for humans to unravel Using probabilistic models and machine learning, MAMs could simulate alternate scenarios, helping to identify "but-for" causation (i.e., whether the harm would have occurred 'but for' the defendant's/States actions). It is asserted here that this insight, objectivism, and reflection could lead to greater IHL compliance and less harm. By evaluating variables and identifying causal links, MAMs could assist in determining the chain of events that legally links a person/States actions to the harm caused, which is often very complex in multi-causal cases.
- Reduction in Ambiguities of Legal Causation: Legal causation often involves normative judgments about whether an act is sufficiently connected to the harm to be held liable. However, should MAMs be involved in decision making, then it is asserted here that

> MAMs could assess precedents, identifying patterns in previous judgments, and consequently apply legal principles to factual scenarios continuously. This is seen here as ensuring IHL principles are consistently and perpetually applied.

AM/MAMs may be embodied in hardware, which they control the motion off, e.g., robot arms, but they could equally be encased in software on a system deciding the outcome. Both instances will be capable of meeting the actus reus requirements, as they are controlling the action directly, e.g., a robot arm or via another interface, such as delivering a dose of medicine. Further, it is the view here that AM/MAMs would likely enhance the precision, reliability, and objectivity of causation determinations, helping legal practitioners address complex causational inquiries with more data-driven insights. This could result in greater identification of non-IHL compliance and thus highlight areas/practices individuals and/or States need to address.

3.2.1.4 Mens Rea

Mens rea (MR) is the required mental element of a crime, split into intention and recklessness. Intention can be direct, meaning it was the persons aim or purpose, as with murder, or indirect (or oblique), which is harder to define and looks at the intended consequence of a person’s actions.[257]

Proving the intent in the context of a non-human legal entity such as a corporation, is more challenging. The perspective is that a psychological state can be attributed to a non-human legal entity such as a corporation. This considers the non-human legal entity as if it is

[257] As in the case of Woollin [1998] UKHL 28; [1998] 3 WLR 382.

competent of holding mental states and can be transposed to AMs. With this approach, non-human legal entities are viewed as having unique values, cultures, ethos or characters that can produce the opportunity of corporate wrongdoing. Nevertheless, there tends to be a human behind the corporate veil and consequently liable (e.g., corporate manslaughter).

Recklessness usually involves a person taking an unjustifiable risk, with awareness of the risk.[258] Humans can have lapses in judgement and take risks that are unreasonable in the circumstances.[259] Indeed, it is said that to err is human, nevertheless, an AM could be in a position whereby it takes an unreasonable risk, especially in the throws of conflict. For example, unreasonableness could be considered where an AM under a duty, decides to ignore a piece of data before taking action, having foreseen the consequences, and was unjustified in taking such risk. It is questioned here how we would set/teach the guardrails of when it is justified to take a risk, as this tends to depend upon the situation, environment, knowledge at the time, etc., in addition, the practicalities of this is yet to be understood. Indeed, cause and effect can be perceived differently by individuals, so it is argued here that that each AM would have a differing view of unreasonableness, due to the environment they developed for and operated within. Further, the effects of a human taking an unjustified risk can result in an emotional response, yet this may not arise in an AM. Humans also take risks due to beliefs, intuition, and mental state, which we are yet to know if AMs will have the capacity for.

[258] After a bit of toing and froing over the test for recklessness, the Cunningham [1957] 2 QB 396 subjective test for recklessness was reasserted by the House of Lords in *R v G* [2004] 1 A.C. 1034.
[259] LexiNexis "Recklessness definition' (LexisNexis, 2024) < https://www.lexisnexis.co.uk/legal/glossary/recklessness#:~:text=In%20essence%2C%20recklessness%20means%20the,Lords%20in%20R%20v%20G.> accessed 4 September 2024.

To note, civil litigation generally requires the mental element of negligence to be proved, meaning the defendant must have failed to understand circumstances or consequences that a reasonable man would understand.[260] This mental element is not found in criminal law except in specific instances; manslaughter, which actually requires 'gross' negligence to be proven; and rape and some other sexual offences. One considers the application of the mental element of negligence in relation to MAMs and the applicability. Indeed, it has been argued here that MAMs can meet the criminal elements of AR and MR, however, the circumstances and consequences, just like recklessness, will depend on the situation, environment, access and quality of the information, and data at the time, something which is a challenge in assessing in humans. It also is dependent on the clarity and quality of the orders, and for which we know the English language can have double means, e.g., my children now say something is 'sick', meaning it is great, but the literal mean of 'sick' is that something is not good, poorly or ill – How will a MAM understand and interpret this nuance? Further, this is likely to place a duty on the State/army to ensure the MAM has constant access to validated (e.g., not fake) data, or the State/army could be liable for the negligence, as it could be argued that the consequences resulting from poor or out of date data, were foreseeable.

AMs, with their very complicated decision-making systems and own goals, will make it unfeasible to ascertain the rational for the AM's actions by interrogating the people programming the AM, as the programmers may not know exactly how AMs work, or the information they have access to. Thus, it appears practical to regard AMs as rational agents and strive to find explanations for their actions, which directs us to consider the best

[260] Vaughan v Menlove (1837) 3 Bing NC 467

interpretation and prediction of AMs behaviour. Therefore, just like a human, the AM's mental state will need to satisfy the mens rea requirement of the crime; Either knowledge, intent, or negligence will need to be proven. Hallevy defines knowledge for machines as "sensory reception of factual data and the understanding of that data."[261] Scholten states, "Robots are capable of attaining and processing such knowledge: a self-driving car will receive information through its sensors about a pedestrian crossing the street, algorithms will process and analyse this data and the car will hit the break."[262] Again, this may place a duty on the user to ensure any retained data is still valid, relevant and acceptable. Storing this data could blend into data protection and security laws and policies, which is a concern, but of which are not the focus of this thesis. Nevertheless, as mentioned in chapter 2, trust is key to AMs acceptance and adoption, thus when we sell an AM, we will want to be sure we remove our personal data before it leaves us, but it is questioned if this will be ethical or practical. When a relationship breaks down between partners, we do not and cannot delete the data they hold about us in their memories, but simply hope and trust they will keep it confidential and be careful and mindful over the memories they do share (even the good ones); why should the same not apply to AMs? These are questions that will need looking at alongside the development of AMs/MAMs.

As already stated, presently it is the human who forms the mens rea and then channels that into the operation of the non-conscious AM. Thus, the AM does not form the criminal intent and meet the requirements of mens rea. However, as the conscious AMs discussed in this thesis have goal-driven behaviour, in that they are able to execute action X to reach goal Y, it

[261] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

[262] Nina Scholten, 'The Robo-Criminal' [2019] Artificial Intelligence & Law (Fastcase) 263.

follows that such an AM is capable of intent and thus form the required MR. Further, their goal-driven behaviour has significant implications when looking at strict liability offenses. Strict liability is liability where MR does not have to be established in relation to the AR elements, although intention, recklessness or some knowledge may still be required for the other elements of the offence. The liability is 'strict' due to defendants being convicted irrespective of being genuinely unaware of the elements which made their acts or omissions criminal. As such, what the defendant intended, knew, or believed, is deemed irrelevant. Arguably, strict liability will be easier to establish for AMs, as the AR is the sole focus, however, care must be taken to ensure the AM did indeed carry out the act freely, and it was not a human using the AM as a tool, which is discussed later in section 3.4.2 below.

In proving intent in humans, we try to establish the purpose for an action or if the person knew, or could have been reasonably expected to know, that the outcome was most likely to occur. Obviously, we cannot delve into a person's mind, so instead the CPS[263] examine the circumstances and cross-examine witnesses about the facts or their observations. Based on the evidence, the CPS try to show the person intended the outcome 'beyond reasonable doubt'. However, without a confession it is difficult to prove the actual state of a defendant's mind. Conversely, with AMs, this is perhaps arguably easier, as we may be able to take a look into the AM's 'black box'[264] and actually 'see' the reason and/or thought process of why an AM undertook an action. There are some crimes that require more insight, for example, race or gender-based crimes, where an element of 'hatred' is required. According to Scholten[265],

[263] Crown Prosecution Service
[264] 'Black 'box' takes its origins from aviation technology and is an accident recorder that is developed to "understand what happened and why, therefore determining how to prevent the scenario from happening again." Airbus Safety, 'What is a black box and how does it work?' (Airbus, 16 May 2024) < https://www.airbus.com/en/newsroom/stories/2024-05-what-is-a-black-box-and-how-does-it-work > accessed 4 September 2024.
[265] Nina Scholten, 'The Robo-Criminal' [2019] Artificial Intelligence & Law (Fastcase) 263

"for that small category of crimes, an AM is most likely not able to satisfy the requirements." Arguably, while AMs will not have lust, they could be taught to hate, due to exposing them to negative data.

Importantly Scholten supports the theory that AMs will be able to be criminally liable, by asserting:

> "Since actus reus and mens rea are the only criteria for criminal liability and those can, in most situations, be satisfied by the highly autonomous robots, this leads to the conclusion that robots can be criminally liable for their acts and punishment can accordingly be imposed for such criminal conduct."[266]

Solaiman[267] disagrees that AMs can commit offences in the true sense and states that Hallevy's models "implicitly deny or plainly overlook the fact that any punishment imposed on a corporation effectively punishes human beings behind it (managers and/or owners)."[268] Furthermore, they advocate punishing the "individuals whose fault, if any, caused the robot's [AM's] malfunction contributing to harm sustained by humans."[269] This is again a very simplistic view of AMs and quite honestly ignores the complexity of the conscious AMs we are developing for the future.

Exploring the non-military specific legal landscape, it is the view here that the current legislation does not sufficiently regulate AM development, nor does it address the complex challenges of conscious AMs, where the requirement of 'intent' or even recklessness in their

[266] Nina Scholten, 'The Robo-Criminal' [2019] Artificial Intelligence & Law (Fastcase) 263

[267] S.M Solaiman, 'Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy' (2017) Artif Intell Law 25, 155–179 (2017) <https://link.springer.com/content/pdf/10.1007/s10506-016-9192-3.pdf > accessed on 17th October 2018

[268] S.M Solaiman, 'Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy' (2017) Artif Intell Law 25, 155–179 (2017) <https://link.springer.com/content/pdf/10.1007/s10506-016-9192-3.pdf > accessed on 17th October 2018.

[269] S.M Solaiman, 'Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy' (2017) Artif Intell Law 25, 155–179 (2017) <https://link.springer.com/content/pdf/10.1007/s10506-016-9192-3.pdf > accessed on 17th October 2018

judgment, could be met, resulting in a future liability gap. There appears to be an underestimation of the pace of development and the impact AMs will have on our lives, how we interact with the world, and each other. However, technology and policy leader CAIS[270] have raised growing concerns, stressing that, “AI risk has emerged as a global priority, ranking alongside pandemics and nuclear war. Despite its importance, AI safety remains remarkably neglected, outpaced by the rapid rate of AI development. Currently, society is ill-prepared to manage the risks from AI.”[271] Ignoring warning like this will create significant liability gaps and risk humans abusing and exploiting AMs and vice versa. What is required is a radical rethink of how we identify and apportion liability. In doing so, some commentators consider apportioning liability at stages throughout AM development, which is explored in the section 3.4. This could also be applicable to the military context.

3.3 Exploring IHL

The scope of international law presides over international relations both in peace times and in times of armed conflict. It encompasses aspects such as the demarcation of “international boundaries, international trade, the law of the sea, air and space law, human rights, protection of the environment, and diplomatic relations.”[272] It also controls the situations in which a State may use armed force (‘jus ad bellum’) and the manner in which armed force is

[270] CAIS, ‘CAIS About’ (CAIS, 2023) <https://www.safe.ai/about > accessed 6 June 2023.
[271] CAIS, ‘CAIS About’ (CAIS, 2023) <https://www.safe.ai/about > accessed 6 June 2023.
[272] Director General, ‘JSP 383: The Joint Service Manual of The Law Of Armed Conflict’ (Government Publishing, 2004) < https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/27874/JSP3832004Edition.pdf > accessed 3 March 2017.

in reality used (‘jus in bello’ or ‘the law of war’). The BA’s JSP 383[273] is focused on jus in bello, which today is more commonly known as the LOAC or IHL. It should be noted that the BA deem that, “the law of armed conflict, being part of international law, is binding on states and also regulates the conduct of individuals.”[274] To stress again, ‘individuals’ here are humans and not any other conscious entity, such as a MAM.

Sources of IHL can be found in:

- Customary law – “rules developed from the practice of states which are binding on states generally.”[275] The rules are a result of State practices observed over a period of time, together with ‘opinio juris’ which is, “a belief on the part of the state concerned that international law obliges it, or gives it a right, to act in a particular way;”[276] and
- Treaty law – “rules expressly agreed upon by states in international treaties which are only binding on states party to those treaties.”[277]

As discussed in chapter 2, IHL is regarded as a set of rules for war and aims to balance legitimate military action with reducing human suffering, especially civilians. To stress, these principles apply to humans and do not make any allowances for MAMs (non-conscious MAMs

[273] Director General, ‘JSP 383: The Joint Service Manual of The Law Of Armed Conflict’ (Government Publishing, 2004) < https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/27874/JSP3832004Edition.pdf > accessed 3 March 2017.

[274] Director General, ‘JSP 383: The Joint Service Manual of The Law Of Armed Conflict’ (Government Publishing, 2004) < https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/27874/JSP3832004Edition.pdf > accessed 3 March 2017.

[275] Director General, ‘JSP 383: The Joint Service Manual of The Law Of Armed Conflict’ (Government Publishing, 2004) < https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/27874/JSP3832004Edition.pdf > accessed 3 March 2017.

[276] Director General, ‘JSP 383: The Joint Service Manual of The Law Of Armed Conflict’ (Government Publishing, 2004) < https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/27874/JSP3832004Edition.pdf > accessed 3 March 2017.

[277] Director General, ‘JSP 383: The Joint Service Manual of The Law Of Armed Conflict’ (Government Publishing, 2004) < https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/27874/JSP3832004Edition.pdf > accessed 3 March 2017.

are rightly under human control). These rules, whilst in the throes of battle and having to make split second decisions, can be challenging for soldiers to interpret and comply with. Yet, there have been instances when soldiers have deliberately stepped outside of these laws, resulting in devastating consequences, such as in the British case of Marine A, who was later disclosed as Alexander Blackman[278]. Nevertheless, it must be stated that cases such as Marine A are thankfully rare and the majority of the British Army (BA) comply with all IHL.

3.4 Implications for AM/MAMs with regards to Liability under Current Legal Framework

3.4.1 AM/MAMs as a Victim, Agent, and Perpetrator

The question of whether criminal law could be applied to humans for the actions of AI as asked by Asaro[279] and Pagallo[280], could be answered by the principle of innocent agency. Thus, if an AM followed the directions of a human and the actions would be a crime if the human undertook it themselves, then the acts of the AM would generally be ascribed to the human; Under the control with the human.[281] This is akin with an adult asking child to commit a crime. Here the principal must have mens rea for the crime.[282]

[278] R v Sergeant Alexander Wayne Blackman ("Marine A"), Case Reference: 2012CM00442. < https://www.judiciary.uk/wp-content/uploads/JCO/Documents/Judgments/r-v-blackman-marine-a-sentencing+remarks.pdf > accessed 3 March 2020.

[279] Peter Asaro, 'The Liability Problem for Autonomous Artificial Agents' [2016] Association for the Advancement of Artificial Intelligence.

[280] Ugo Pagallo, The Laws of Robots: Crimes, Contracts, and Torts (2013 edn, Springer).

[281] P Alldridge, 'Crim Law Forum ' (1990) 2: 45 Williams, G. Crim Law Forum (1992) 3: 289 < https://doi.org/10.1007/BF01096228 > accessed 10 November 2017.

[282] R v Jogee [2016] UKSC 8.

In Jogee[283], the court held that to establish "accessorial liability"[284], the element of conduct must be proved in addition to the required mental element. This could be satisfied by evidencing that the accessory either assisted or, as a minimum, encouraged the principal in committing the offence. The mental element is satisfied by proving the accessory intended to assist or encourage the principal, however, the mental element is not satisfied by simple foresight that the principal may commit an offence,[285] consequently, it is the view here that this would be of great interest to manufacturers who did not have any intention. Moreover, drawing from Jogee, the author identifies the following facets for this assumed interest:

- Accessory Liability and Intention to Assist or Encourage: If a manufacturer supplies tools or devices knowing they will be used for illegal activities, then they may be liable as an accessory. [286] Thus, for a manufacturer of AM/MAM, a concern arises if their product decides to act in a criminal way. The manufacturer may worry about being held liable if it is proven the AM/MAM intended to assist or encourage a crime, via the design or programming of the machine, in such a way that it facilitated a criminal act(s).
- Lack of Criminal Intent by the Manufacturer: It is assumed here that the majority, if not all, manufacturers do not intend for their machines to commit crimes. However, AM/MAMs will act on reasoning processes independent of the manufacturer's intentions and in environments (e.g., war) that are outside of their scope of control. This disconnect is argued here as challenging for courts to apply accessory liability,

[283] R v Jogee [2016] UKSC 8.

[284] In the case of *R v Jogee; Ruddock v The Queen* [2016] UKSC 8; UKPC 7 the Supreme Court and the Privy Council focused the controversial principle of "parasitic accessory liability" (PAL). PAL applies in a situation where two people (D1 and D2) set out to commit crime A, and in the course of that venture D1 commits crime B, D2 would be guilty as an accessory to crime B if he had foreseen the possibility that D1 might act as he did.

[285] R v Jogee [2016] UKSC 8.

[286] CPS, 'Secondary Liability: charging decisions on principals and accessories' (CPS, Revised: 04 February 2019, 28 November 2023, 4 July 2024) < https://www.cps.gov.uk/legal-guidance/secondary-liability-charging-decisions-principals-and-accessories > accessed 4 September 2024.

due to the manufacturer lacking intent, or even foreseeing the outcome. Further, it is expected here that manufacturers would be concerned about liability for crimes AM/MAMs may decide to commit. This raises questions about how the law should assess the mental element when human actors (e.g., the manufacturer) have no criminal intent but their AM/MAMs make independent decisions that lead to criminal outcomes.

- Strict Liability Risks for Manufacturers: As mentioned, under English law, strict liability is imposed on manufacturers in certain contexts, thus they could be held responsible without a specific mental element (such as intent). This principle, if extended to AM/MAMs, may result in manufacturers facing liability purely because they created a AM/MAM capable of causing harm, despite not intending or foreseeing the specific criminal outcome.
- The Risk of Being Held Liable for 'Foreseeable' Crimes: Accessories may be liable if they foresee the risk that their assistance/support could encourage a crime, even if it was not their direct intention. For AM/MAMs, manufacturers may worry that courts could interpret foreseeable misuse as grounds for liability. Further, it is the view here that manufacturers are highly likely to be extremely concerned where MAMs are involved, due to the uniqueness of the military environment. Thus, it is argued here as unjust to expose manufacturers to liability for crimes, even war crimes, regardless if they could have arguably foreseen it, but where MAMs will have the capacity for autonomous decision-making.
- Legal and Ethical Implications for Manufacturers: If AM/MAMs are developed and deployed with full autonomy (discussed in chapter 4), does that autonomy include potential criminal behaviour, including war crimes? Indeed, would it be ethical or even

legally permissible to constrain AM/MAMs actions too strictly, thus impeding their autonomy? These considerations add complexity to both the manufacturing and regulation of AM/MAMs in the legal system.

- Vicarious Liability Challenges: Finally, and aside, AM/MAMs, acting independently, will distort the traditional lines of liability, raising the question of whether manufacturers should be treated as vicariously liable for the AM/MAMs actions. Traditionally, vicarious liability applies to employers for their employees' actions, but applying this to AM/MAMs would be novel, complex, and contentious. However, whilst of interest, this area of employment law is not discussed further.

It is asserted here that the reasons above highlight that manufacturers of AM/MAMs would likely be highly concerned about liability for crimes, including war crimes, which their AM/MAMs may commit, especially given the complexities in proving intent and foreseeability. However, a potential solution and beacon of hope to such liability challenges, is offered up by Hallevy.

Hallevy has dedicated much attention to the subject of AM criminal control and responsibility, and, at the time of writing, is the only person who is forward thinking enough to proposes three models that are formed around AMs. It is the view here that his models take a practical and pragmatic approach, and could be tailored to fit within legal systems, such as the English legal system. As introduced in chapter 2, his three models are; 1) the perpetration-by-another responsibility model, 2) the natural-probable-consequence responsibility model and, 3) the direct responsibility model.[287] He views these three models he views as a solution to the

[287] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

problem of AMs meeting the actus reus and mens rea requirements for criminal responsibility. Hallevy's models are considered here as a continuation of current extrinsic control[288].

Hallevy envisions that the three models may be employed separately, but that various situations will require "a coordinated combination of them (all or some of them)… in order to complete the legal structure of criminal responsibility."[289] Hallevy's models are supported by Scholten.[290] Despite Solaiman[291] supporting the concept of Hallevy's models, he does not see future where AMs will be able to have actus reus and mens rea through their own actions, consequently does not see the models as having any real-world application. Indeed, it is this short-sightedness that risks widening the liability gap and hinders evolvement of our legal system.

Hallevy observes that his models provoke questions over what is the "essence of humanity (Do human beings function as thinking machines?)"[292] and of AMs "Can there be thinking machines?",[293] and proposes five qualities that an intelligent entity should have:

> (1) Communication: "One can communicate with an intelligent entity. The easier it is to communicate with an entity, the more intelligent the entity seems. One can

[288]

[289] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

[290] Nina Scholten, 'The Robo-Criminal' [2019] Artificial Intelligence & Law (Fastcase) 263.

[291] S.M Solaiman, 'Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy' (2017) Artif Intell Law 25, 155–179 (2017) <https://link.springer.com/content/pdf/10.1007/s10506-016-9192-3.pdf > accessed on 17th October 2018.

[292] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

[293] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

communicate with a dog, but not about Einstein's theory of relativity. One can communicate with a little child about Einstein's theory, but it requires a discussion in terms that a child can comprehend."[294]

(2) Internal Knowledge: "An intelligent entity is expected to have some knowledge about itself."[295]

(3) External Knowledge: "An intelligent entity is expected to know about the outside world, to learn about it, and utilize that information."[296]

(4) Goal-Driven Behaviour: "An intelligent entity is expected to take action in order to achieve its goals."[297]

(5) Creativity: "An intelligent entity is expected to have some degree of creativity. In this context, creativity means the ability to take alternate action when the initial action fails. A fly that tries to exit a room and bumps into a windowpane, tries to do that over and over again. When an AI robot bumps into a window, it tries to exit using the door. Most AI entities possess these five attributes by definition."[298]

Meeting the above five qualities will involve highly sophisticated architecture, that will integrate self-awareness, environmental perception, adaptive goal setting, creative processing, and contextual communication, which is discussed in detail in chapter 5. Consciousness would likely emerge from the synergy between these five qualities, all

[294] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

[295] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

[296] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

[297] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

[298] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

dynamically interacting in real-time, thus not from any single component. This integrated design would allow AM/MAMs to interact intelligently and flexibly with humans and their environment in ways that feel purposeful, responsive, and creative. To reiterate, the AMs, and subsequently MAMs, discussed in this thesis possess all 5 of the above qualities, which is where the liability challenges emerge.

3.4.2 Hallevy's Models for AM/MAM Liability

Exploring each of Hallevy's models in depth, we start with the "Perpetration-by-Another Virtual Responsibility Model".[299] This model does not regard an AM as having any human qualities and the AM is deemed an innocent agent; "a machine is a machine, and is never human."[300] Thus, the AM does not have sufficient capability to be considered the perpetrator of an offence and the programmer is the perpetrator. Hallevy views this as resembling "the parallel capabilities of a mentally limited person, such as a child, or of a person who is mentally incompetent or who lacks a criminal state of mind."[301]

Hallevy identifies a potential second "perpetrator-via-another" as being the user of the AM, who exploits the AM for their own gain, but who did not actually program the software. Hallevy illustrates this with a hypothetical robot example:

> "…a user purchases a servant-robot, which is designed to execute any order given by its master. The specific user is identified by the robot as that master, and the master orders the robot to assault any invader of the house. The robot executes the order exactly as ordered. This is not different than a person who orders his dog to attack any

[299] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.
[300] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.
[301] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

> trespasser. The robot committed the assault, but the user is deemed the perpetrator."[302]

It was the AM that committed the offense in both the above situations, relieving the programmer or user of the actus reus requirement of the specific offense. However, this would undoubtedly prove unacceptable and create a loophole to be exploited. Police and/or guard dogs have to comply with strict policies.[303] Ultimately it is the human handler that is responsible for the harm caused, if acting under the direction of a police officer, and the dog is "deployed as an instrument of force."[304] This categorises these types of dogs more as tools than autonomous entities. This fits the current situation, with military machines and their limited autonomy; the human would be liable for breaches of IHL. Hallevy considers the Perpetration-by-Another responsibility model as a solution, which judges the actions executed by an AM;

> "As if it had been the programmer's or the user's action. The legal basis for that is the instrumental usage of the AI entity as an innocent agent. No mental attribute required for the imposition of criminal responsibility is attributed to the AI entity. When programmers or users use an AI entity instrumentally, the commission of an offense by the AI entity is attributed to them."

Further, this model avoids assigning mental capability (including human mental capability) to the AM and instead presumes mens rea of the programmers or users. Hallvey clarifies by saying:

> "According to this model, there is no legal difference between an AI entity and a screwdriver or an animal. When a burglar uses a screwdriver in order to open up a

[302] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

[303] West Yorkshire Police, 'Police Dog Bites' (West Yorkshire Police, 2020) < https://www.westyorkshire.police.uk/sites/default/files/2020-06/police_dog_bites.pdf > accessed 20 June 2020.

[304] Alliance Dogs Unit, Dog Legislation Officer (14447), *J-P-063* (Devon and Cornwall Police, 11 January 2022).

> window, he uses the screwdriver instrumentally, and the screwdriver is not criminally liable."[305]

This model guards against the potential exploitation of AMs and using them to avoid liability. This first model could be implemented immediately for protection against offences, as it matches the maturity of today's AMs.

The second of Hallevy's models is the "Natural-Probable-Consequence Virtual Responsibility Model." This model presumes significant participation of the users or programmers in the AMs day to day tasks, although they have no intention of committing an offence through the AM.

Hallevy illustrates this second model via scenarios, with the first being:

> "During the execution of its daily tasks, an AI entity commits an offense. The programmers or users had no knowledge of the offense until it had already been committed; they did not plan to commit any offense, and they did not participate in any part of the commission of that specific offense."[306]

He follows this with an illustrative example in which an AI, functioning as an automatic pilot, overrides a human pilot's commands to protect its mission, resulting in lethal consequences for the pilot. This example emphasises the potential autonomy and unforeseen actions of AI systems during operation.[307]

[305] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

[306] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

[307] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

It is clear that the programmer did not intend for anyone to be harmed, let alone killed, however unfortunately the pilot was killed through the AI's action, which were in accordance with the software programme. It is for situations such as the pilot example above that Hallevy advocates using his "natural probable consequence responsibility" model, saying it is:

> "Based upon the ability of the programmers or users to foresee the potential commission of offenses… Natural-probable-consequence responsibility seems to be legally suitable for situations in which an AI entity committed an offense, while the programmer or user had no knowledge of it, had not intended it and had not participated in it."[308]

It is the view here that Hallevy's second model could be adapted and applied to criminal law. It could be argued under English law that the user or programmer could have oblique intention, in that they knew the outcome was almost a certainty or was reckless and took un unjustifiable risk. Thus, for criminal liability to be held, all that this model needs to be adapted for is recklessness (under English law) on the part of the programmer or user, along with the degree of "natural probable consequence", which is in English law known as foreseeability:

> "The natural-probable-consequence responsibility model would permit responsibility to be predicated upon negligence [recklessness under English law], even when the specific offense requires a different state of mind. This is not valid in relation to the person who personally committed the offense, but rather, is considered valid in relation to the person who was not the actual perpetrator of the offense, but was one of its intellectual perpetrators. Reasonable programmers or users should have foreseen the offense, and prevented it from being committed by the AI entity."[309]

This model has the appeal that recklessness does not have as high a threshold as intention. It seems to offer a good framework for recourse in situations where it would be unjust to

[308] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

[309] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

presume intention on the part of the programmer or user but allows for liability for acts or omission which would fail to meet the 'virtually certain' standard of Woollin for intention.[310]

Hallevy's third model is the "Direct Virtual Responsibility Model", which, "does not assume any dependence of the AI entity on a specific programmer or user,"[311] but concentrates solely on the AM itself, which would make Broozek and Jakubiec shiver![312] Here the AM must be proven to have the mens rea and actus reus to be held criminally liable. This requires an AM to confidently, and clearly, satisfy the conditions for criminal liability and raises a question around if AMs are to be treated differently in this situation. Indeed, Hallevy highlights that just because AMs will have many features and capabilities that will surpass humans, these should not elevate or hasten the assignment of criminal liability. He emphasises that when an entity satisfies the requirements of mens rea and actus reus, then criminal liability is enforced, saying:

> "If an AI entity is capable of fulfilling the requirements of both the external element and the internal element, and, in fact, it actually fulfils them, there is nothing to prevent criminal responsibility from being imposed on that AI entity."[313]

As mentioned, usually it would be clear if the AM satisfied the actus reus, providing that the AM controls its moving parts, methods or tools to perform the action, for example a mechanical arm. Hallevy gives the example of assault, whereby an AM moves its mechanical arm and hits a person close by, easily satisfying the condition of actus reus.

[310] Woollin [1998] UKHL 28; [1998] 3 WLR 382.

[311] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

[312] Bartosz Broozek and Marek Jakubiec, 'On the legal responsibility of autonomous machines' [2017] Artif Intell Law (2017) 25:293-304

[313] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

Where an AM is under a duty to act, yet fails to act, then this satisfies the actus reus of an offence. Therefore, just like any human or organisation, an AM's omission will represent the actus reus and establish liability for offences where the law imposes a duty to act and the AM is in breach of that duty.

The third model is of most interest in this thesis and guards against the future conscious AMs and allows for AMs to be assigned criminal responsibility. Thus, under this model, AMs will have rights and duties. In addition, Hallevy contemplates the challenge around an AM using another AM to commit a crime and recommends his third model as a remedy to this, along with the first model (he states his models can be combined as appropriate), saying:

> "The third responsibility model in that [aforementioned] situation is applied in addition to the first responsibility model, and not in lieu thereof. Thus, in such situations, the AI entity programmer shall be criminally liable, pursuant to a combination of the Perpetration-by-Another responsibility model and the direct responsibility model. If the AI entity plays the role of the physical perpetrator of the specific offense, but that very offense was not planned to be perpetrated, then the application of the natural-probable-consequence responsibility model might be appropriate… when the programmer is not human, the direct responsibility model must be applied in addition to the simultaneous application of the natural-probable consequence responsibility model; likewise, when the physical perpetrator is human while the planner is an AI entity."[314]

Overall, Hallevy's first two models provide an effective framework for identifying and assigning criminal liability, that looks to address the potential problem of people committing crimes via an AM and expecting to avoid liability. It is the view here that these two models should be implemented immediately to close gaps and in readiness for the legal situations they aim to address. All three models appear to be able to handle, and cater for, AMs as they develop and become legal persons, have rights of their own, and even consciousness. For

[314] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

example, the first model could be used by an AM if it felt it was being made to do something under duress by a human. It looks to protect the AM from abuse and coercion, much as it would a child or vulnerable person. Hallevy's idea for combining the models for situations where an AM uses another AM to commit a crime, again guards against abuses of power and duress.

The third model is viewed here as an effective model to address the legal challenges of conscious AMs. The third model can be implemented when we near the edges of AM consciousness, ensuring we are prepared and will have legal certainty established and embedded. Scolten[315] further analyses Hallvey's model and affirms that the actus reus and mens rea can be satisfied by AMs and even agrees with Hallvey that AMs can be held criminally liable and, arguably more significant, effectively punished for criminal acts. However, the models have yet to be tested, so there may be issues with application in the real world, nevertheless, based on thought experiments, they offer a reasonable starting point and protection.

Hallevy's critics such as Solaiman[316] see Hallevy as a pioneer in robot (AM) criminal responsibility and liability. Despite the criticism and likely muffled laughter of Hallevy's futuristic approach to AMs, critics such as Soliaman, do see value, logic and theoretical applicability in his model, even though they do not conceive a day when AMs will ever be that advanced and out of the control of humans.

[315] Nina Scholten, 'The Robo-Criminal' [2019] Artificial Intelligence & Law (Fastcase) 263.
[316] S.M Solaiman, 'Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy' (2017) Artif Intell Law 25, 155–179 (2017) <https://link.springer.com/content/pdf/10.1007/s10506-016-9192-3.pdf > accessed on 17th October 2018.

Humans do not take comfort in having liability gaps and need to find and punish the person responsible for harm. Therefore, how we will handle and accept an AM being criminally responsible is still unknown. Nevertheless, we even call natural disasters ‘Acts of God’ because we feel there should be someone to blame, divine or otherwise. Consequently, we may require a gradual processes of criminal liability transference to AMs, so our instincts, attitudes, and legislation can become acclimatised. It is the view here that a gradual process of liability transference will provide comfort, build trust, and test the process. It will also allow for reviews and adaptions, even to stop the transference, if required.

3.5 Application of MAMs to Currently Legal Landscape

3.5.1 Understanding the Impact

This section of the chapter takes the learning from criminal liability, in that AM/MAMs will be fully conscious and, as such, able to meet the mens rea and actus reus requirements, and considers this against the IHL landscape. It pertains to all other research aims, as the ethical challenges have directly translated into IHL, and State policies. This chapter is only concerned with IHL of which the UK is a signatory[317].

Further, this section concentrates on the potential unique responsibility owed to MAMs, under IHL, which is asserted here as the true value alignment problem (TVAP). Indeed, it is argued here that the TVAP concerns how we recognise, accommodate, and protect the rights

[317] International Humanitarian Law Databases, 'Treaties and States Parties’ (*ICRC,* 2024) < https://ihl-databases.icrc.org/en/ihl-treaties/treaties-and-states-parties > accessed 4 September 2024.

and views of MAMs under IHL. However, it is sadly understood by the author that humans experience challenges in recognising, accommodating, and protecting the rights of other humans (for example the recent Hamas attacks), so extending this to another entity is not only ambitious, but fraught with political and humanitarian rights difficulties. Nevertheless, that should not stop us considering how we realign our vales to accommodate the rights and duties MAMs will owe us, and that we will owe them.

Extrinsic control and criminal cases against the BA do not arise in the same way as for an organisation. States are punished under International Human Rights laws for offenses such as war crimes, however, individual members of a UN signatory army can be convicted in the criminal court for unauthorised action, which demonstrates why criminal liability was previously explored, and will be discussed later. In addition, civil cases are used to highlight the BA's responsibility towards human soldiers, which should be considered and applied to MAMs.

MAMs introduce novel challenges into the battlefield, such as programming them to obey the core principles IHL of humanity, distinction, proportionality, and necessity, for which the application and applicability for MAMs, is discussed in chapter 5. Accordingly, the technology must be controlled and monitored, and we must not underestimate the impact they will have. So fearful are some experts, that they have called for the UN to ban certain autonomous weapons under the Convention on Certain Conventional Weapons (CCW).[318] This would ban

[318] 51 states negotiated the Convention on Certain Conventional Weapons (CCW) in 1980, which is formally known as the 'Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects' (also referred to as the Inhumane Weapons Convention). It aims to

or restrict “the use of specific types of weapons that are considered to cause unnecessary or unjustifiable suffering to combatants or to affect civilians indiscriminately.”[319] However, in 2015 the UK disputed such a ban stating:

> “At present, we [the Foreign Office] do not see the need for a prohibition on the use of Laws, as international humanitarian law already provides sufficient regulation for this area. The United Kingdom is not developing lethal autonomous weapons systems, and the operation of weapons systems by the UK armed forces will always be under human oversight and control.”[320]

Subsequently, if MAMs have a form of legal personhood, then they too should be afforded protection and not subjected to unnecessary harm and suffering.[321] This would impose a duty (or burden) on the BA to assess the impact and risk of operations to MAMs, just as it does for human soldiers. Indeed, the consideration and application of IHL principles to MAMs is argued here as a real challenge to our IHL values and will require us to realign them to a non-human, conscious, entity; The TVAP, and which is the focus of chapter 6.

3.5.2 Upholding IHL Challenges

Marine A was convicted of murder on September 15, 2011, for shooting an injured Afghan insurgent who was no longer a threat after being wounded by an Apache helicopter. During the incident, Marine A moved the insurgent out of surveillance view, denied him medical aid,

protect military troops from inhumane injuries and prevent non-combatants from accidentally being wounded or killed by certain types of arms. (https://www.armscontrol.org/factsheets/CCW).

[319] United Nations, Office for Disarmament Affairs, ‘The Convention on Certain Conventional Weapons’ (United Nations, 2017) < https://www.unog.ch/80256EE600585943/(httpPages)/4F0DEF093B4860B4C1257180004B1B30?OpenDocument > accessed on 17 November 2017.

[320] Owen Bowcott, UK opposes international ban on developing 'killer robots' (The Guardian, 13 April 2015) < https://www.theguardian.com/politics/2015/apr/13/uk-opposes-international-ban-on-developing-killer-robots > accessed 10 November 2017.

[321] AM suffering and harm is discussed in the Ethics chapter, Part I.

and then shot him at close range while acknowledging his actions violated the Geneva Convention. The court, composed of experienced military personnel, stressed the need for British Armed Forces to maintain the highest standards of humanitarian conduct to avoid comparisons with insurgent behaviour and retain global trust.

Marine A was sentenced to life imprisonment with a minimum of ten years, but this was later reduced to manslaughter due to diminished responsibility on appeal, as he was shown to be suffering post-traumatic stress disorder (PTSD).It is the view here that in such high-pressure situations, MAMs may outperform human soldiers, as they will not suffer from stress or PTSD, in conjunction with maintaining consistent and auditable decision-making. However, this highlights the necessity of ensuring MAMs are developed and trained to adhere to International IHL and uphold the values to avoid liability issues.

With regards to the BA, soldiers are trained, and then observed, to ensure they take orders from their seniors, and act in accordance with those orders. This is termed Mission Command. Mission Command is the:

> “Practice in the UK's Armed Forces of devolving responsibility down to low levels of command… The commander's intent is shared with subordinates, who are told what to achieve and why, but are then left to decide how to achieve it. Subordinates are encouraged to use their judgement, initiative and intelligence in pursuit of the commander's goal.”[322]

Mission command is not just the BA’s philosophy of command, it is in actuality the NATO philosophy of command, yet again, with the assumption of a human soldier in mind. Founded on clear expression of intent by commanders, it allows subordinates the freedom to act to

[322] Parliament, ‘5 Command Issues: Mission Command’ (2004) Parliament.uk <https://publications.parliament.uk/pa/cm200304/cmselect/cmdfence/465/46508.htm> accessed 20 February 2017.

achieve that intent, ever mindful of the strict rules they operate within. It is this 'freedom to act' that is of most interest with MAMs along with their freedom not to act (Owen's 'free won't'). It is also the key area where the 'true' value alignment problem (TVAP) becomes glaring obvious; IHL assumes those fighting as soldiers do so with consent and through their own choice. MAMs will not consent but be designed to be deployed. As a result, it is argued here that this is the pinnacle of the IHL TVAP, due to the following aspects:

- Consent: Unlike human soldiers who choose to join the military and consent to all it entails, it is the view here that MAMs will not be asked to consent to participation and will instead be deployed without choice. This leads to ethical dilemmas about coercion and potential violation of their 'rights', which is explored in detail in chapter 6.
- The Absence of MAM Agency: IHL assumes human combatants have agency and are capable of judgment and accountability, with 'accountability following control'[323]. MAMs currently lack agency due to their programmed nature, thus raising questions about consent and legitimacy, and which is analysed in chapters 5 and 6.
- Accountability and Responsibility: Assigning responsibility for actions made by MAMs is complex, and can involve a combination of the MAM, its manufacturer/developer, and military commanders, which disrupts traditional command structures. Decision making is explored further in chapter 5.

In addition, to the TVAP aspects above, the MoD may one day soon owe a duty of care to MAMs. Indeed, the MoD also has a duty of care towards its soldiers, as determined in the

[323] Ministry of Defence, JSP 815. Element 5: Supervision, Contracting and Control Activities (JSP 815) Ministry of Defence < https://assets.publishing.service.gov.uk/media/66e18438dd4e6b59f0cb2500/JSP_815__Element_5_Supervision__contracting_and_control_activities_v1.2.pdf > accessed 4 September 2024.

case of Smith, also known as the Snatch Land Rover Case.[324] This was a civil case and used here to highlight the responsibility owed to soldiers.

It was claimed that the MoD breached its obligation to safeguard life protected by ECHR art 2[325], in that it failed to take, "preventive measures to protect life in the light of the real and immediate risk to life of soldiers who were required to patrol in Snatch Land Rovers."[326] The case, which encompassed claims involving human rights[327], and negligence, also touched on IHL principles, particularly those related to the protection of individuals in armed conflict and the obligations of the British State towards their military personnel.

When analysing the case in relation to IHL, the following is of importance here:

- IHL aims to limit the effects of armed conflict on soldiers and civilians by ensuring humane treatment and minimising unnecessary suffering. While IHL primarily regulates the conduct of hostilities and the treatment of individuals during warfare, this case explores the duty of care states owe to their soldiers, emphasising obligations that align with IHL's protective nature. The MoD's responsibilities for adequately equipping and training soldiers reflect principles found in IHL, such as ensuring proportionality, distinction, and precautions during military operations to protect against harm.

[324] Smith and others (Appellants) v The Ministry of Defence (Respondent) Ellis (Respondent) v The Ministry of Defence (Appellant) Allbutt and others (Respondents) v The Ministry of Defence (Appellant) [2013] UKSC 41.
House of Lords Constitution Committee, Second Report, 2013-4 Session, "Constitutional arrangements for the use of armed force", para 55.

[325] Art 2 regards the right to life and provides that the State should safeguard life and take measures to investigate a death. Smith confirmed that the right to life secures the responsibility of the British government for the deaths of soldiers in combat, killed by enemy troops or illness, if their death is due to inadequate equipment or medical provisions/care. If forces serving abroad are not within the State's jurisdiction under Art 1 then the duties under Art 2 do not apply.

[326] Smith and others (Appellants) v The Ministry of Defence (Respondent) Ellis (Respondent) v The Ministry of Defence (Appellant) Allbutt and others (Respondents) v The Ministry of Defence (Appellant) [2013] UKSC 41.

[327] Article 1 and 2 of the European Convention on Human Rights. Article 1 of the European Convention on Human Rights provides that rights and freedoms should be available to all those within the State's jurisdiction.

- The Supreme Court's decision[328] to allow the negligence claims to proceed signals an extension of scrutiny regarding military conduct and state obligations under human rights law, which can complement and interact with IHL, reinforcing a duty to protect life under various scenarios, even for combatants.

When applying the Smith[329] case to the Value Alignment Problem, the author stresses the following:

- The value alignment problem 'traditionally'[330] refers to aligning systems or actions with human values to ensure safety and ethical conduct. In this context, the MoD's failure to provide proper equipment or training, highlights the challenges of aligning institutional actions with core values, such as ensuring reasonable protection for military personnel, and the safeguarding of their life.
- The judgment raises questions about whether military institutions sufficiently prioritise the protection of their personnel in terms of both legal and ethical values. It reflects broader concerns in governance and military decision, regarding the tension between balancing operational strategic outcomes and the ethical imperative to minimise risks to soldiers' lives. This echoes the challenges of aligning actions with overarching human values.

[328] Smith and others (Appellants) v The Ministry of Defence (Respondent) Ellis (Respondent) v The Ministry of Defence (Appellant) Allbutt and others (Respondents) v The Ministry of Defence (Appellant) [2013] UKSC 41.

[329] Smith and others (Appellants) v The Ministry of Defence (Respondent) Ellis (Respondent) v The Ministry of Defence (Appellant) Allbutt and others (Respondents) v The Ministry of Defence (Appellant) [2013] UKSC 41.

[330] The author proposes a revised term; the true value alignment problem (TVAP), which recognises the interests of MAMs.

Drawing from this case and the analysis, the author is aware that MAMs could introduce vulnerabilities for human soldiers if, for example, their software is out of data, their hardware does not comply with the latest standard, or the MAM fails to pass conformity checks. On the flip side, it is raised here that a MAM that knows it is operating with vulnerabilities which have been overlooked or ignored by the MoD, perhaps could sue the MoD for negligence and for being reckless in their duty to prevent harm to the MAM.

The BA are looking at how to best adapt mission command to meet the demands of digital/information age. The BA see mission command as capturing one of the best facets of the British attitude towards the use of military force, which is, "the ability for a commander to articulate his intent and for the people beneath him to decide on the best way of carrying that out."[331] The digital age should enable better dissemination and explanation of the commander's intent. Nonetheless, there is a risk that mission command in itself can foster a goal-focused (commander's intent) fixation rather than a focus on the outcomes within the new operational environment, which could be challenged by the acts of those in the tactical ranks. Thus, more Junior ranks turn out to be strategically important. This is a potential area where MAMs could be exploited and with, or out of sight of, the oversight of the commanding officers.

The case also shows an "increasing willingness of the courts to become involved in decisions relating to the battlefield," which leads to concern for the MoD and individual military

[331] Parliament, '5 Command Issues: Mission Command' (2004) Parliament.uk <https://publications.parliament.uk/pa/cm200304/cmselect/cmdfence/465/46508.htm> accessed 20 February 2017.

personal. Indeed, the Constitution Committees report[332], looking at use of armed force, agreed that the "negative effect on the morale and operational independence of the armed forces"[333] was a valid concern about "courts scrutinising operational decisions"[334] The main objection is that it hinders the commander's freedom to control the battlefield and respond to intelligence as they see fit. Morgan states that the:

> "Refusal of the Supreme Court in Smith to strike out the claims brings a real risk of defensive decision-making among military planners and commanders. Moreover, it inevitably requires judicial examination of sensitive matters of high national policy."[335]

Whilst it is the view here that the battlefield is a complex and fast-moving environment, that does not typically align itself to non-military court jurisdiction, the findings of the Supreme Court for British soldiers is welcomed, in that it reminds the MoD of their duty. The gap that exists for MAMs, such there will be a requirement that MAMs are battlefield ready, so not putting British soldiers at risk, will need addressing. This could also open the door for protecting conscious MAMs in action, which is where the IHL true value alignment problem (TVAP) lies.

[332] House of Lords Constitution Committee, Second Report, 2013-4 Session, "Constitutional arrangements for the use of armed force", para 55.

[333] House of Lords Constitution Committee, Second Report, 2013-4 Session, "Constitutional arrangements for the use of armed force", para 55.

[334] House of Lords Constitution Committee, Second Report, 2013-4 Session, "Constitutional arrangements for the use of armed force", para 55.

[335] Dr. Jonathan Morgan, 'Military Negligence: Reforming Tort Liability after Smith v. Ministry of Defence Paper presented to the House of Commons Defence Select Committee' [November 2013] Corpus Christi College, University of Cambridge < https://www.biicl.org/files/6759_military_negligence_paper-_jonathan_morgan.pdf > accessed 10 March 2020.

3.5.3 Autonomous Weapon Systems (AWS) Developed and Deployed under IHL

Flowing from the IHL principles and governance, is the development and use of AWS, which include unconscious MAMs, but notably ignores conscious MAMs, in armed conflict. AWS are a step along the path towards conscious MAMs, however, AWS technology is already in existence (discussed in chapter XX), therefore very much a risk of today. The governance includes the duty to carry out legal reviews in the development, acquisition, and implementation of new weapons, as mandated by Article 36 of Additional Protocol I to the Geneva Conventions[336], which is examined at section 3.5.3.2 below. In addition, this links to the duty of care owed by the BA following the Snatch Land Rover case, as discussed earlier.

When looking at what a 'fully autonomous weapon system' may need to demonstrate in order to comply with IHL, numerous speakers at the 2014 Geneva conference stressed that, "qualitative decision-making is typically required when applying the IHL rules of distinction, proportionality and precautions in attack."[337] For example, "the IHL rule of distinction requires that attacks only be directed at combatants and military objectives. Civilians are protected from direct attack, unless and for such time as they are directly participating in hostilities."[338] The definition of military objectives in found in Additional Protocol I, which defies them as:

> "Those objects which by their nature, location, purpose or use make an effective contribution to military action and whose total or partial destruction, capture or

[336] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

[337] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

[338] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

> neutralization, in the circumstances ruling at the time, offers a definite military advantage."[339]

The MoD, despite not using AWS, has built this into its governance. The requitements of the legal review process form a vital part of the control of the battlefield and upholding IHL, creating a 'cradle to grave' governance. This aims to reduce, or even remove, the challenges around accountability and proportionality.

3.5.3.1 UK Military and AWS Status

The MoD submitted evidence to the House of Commons Defence Committee inquiry in September 2013, which was incorporated into their publication[340], and states that, "no planned offensive systems are to have the capability to prosecute targets without involving a human."[341] The MoD made it clear that for current automated weapon systems, human control could be understood as the human deciding and in-putting the pre-programmed parameters of the weapon system's operation. Looking at this from a UK legal perspective, all weapons/machines developed or acquired are subject to legal review in accordance with Article 36 of 1977 Additional Protocol 1 to the Geneva Conventions.[342] These legal reviews include:

[339] Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflict (Additional Protocol I or AP I) (adopted on 8 June 1977, entered into force on 7 December 1978), art 52(2).

[340] House of Commons Defence Committee, 'Remote Control: Remotely Piloted Air Systems - current and future UK use' (2014) House of Commons Sixth Special Report of Session 2014–15 < https://publications.parliament.uk/pa/cm201415/cmselect/cmdfence/611/611.pdf > accessed 17 November 2013

[341] House of Parliament, Defence Committee, 'Written evidence from the Ministry of Defence' (Parliament.co.uk, 2013) < 'https://publications.parliament.uk/pa/cm201314/cmselect/cmdfence/772/772vw02.htm > accessed 10 November 2017

[342] Ministry of Defence, 'UK weapon reviews' (2016) <https://www.gov.uk/government/publications/uk-weapon-reviews > accessed 10 November 2017.

> "an assessment of the compatibility of the weapon with the core rules of IHL[343] as well as an assessment of whether the weapon is likely to be affected by the current and future trends in the development of IHL. The UK considers the existing provisions of international law sufficient to regulate the use of autonomous weapon systems."[344]

At the ICRC, speakers[345] affirmed that the current UK policy states that the "autonomous release of weapons" are not allowed and that, "operation of weapon systems will always be under human control."[346] Accordingly, it is a matter of policy that the UK will, for the near future, remain committed to employing remotely piloted, instead of highly automated systems, to ensure absolute control and authority for weapons release, which provides a degree of comfort. Thus, accountability lies with a human operating the AWS; 'Accountability follows control'. However, as cited, MAMs are not considered here, and therefore no concern raised or theorised over the legal risks and challenges they will introduce. This is regarded as satisfactory if only non-conscious MAMs will be deployed, but stirs a deep anxiety that there is an important oversight with regards to conscious MAMs, and the risks they will introduce to the safeguarding and upholding of the IHL principles.

[343] International Human Rights Law.

[344] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

345 ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

[346] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

3.5.3.2 Ensuring Legal Compliance

Legal reviews form an important part of ensuring compliance to the Geneva Convention[347], and aims to determine "whether new weapons, means or methods of warfare may be employed lawfully under International Law."[348]

Legal reviews of new autonomous weapons should be undertaken throughout the development lifecycle, and not just before taking delivery of the weapon, as there is an interest in safeguarding that the weapon complies with IHL before significant development investment.[349] The speakers at the 2014 Geneva conference were concerned about how the changeable degrees of unpredictability would be tested. This was met with mixed responses and levels of comfort. To reiterate, this thesis does not focus on the technical development of MAMs, however, it is worth noting relevant challenges and this is routinely undertaken by the MoD.[350]

All of the speakers recognised the vast difficulty of the assessments and judgements in applying the:

> "IHL rules of distinction, proportionality and precautions in attack, especially in dynamic conflict environments. These assessments and judgements appear to be uniquely human (some referred to "subjective" appreciation), and would seem extremely challenging to program into an autonomous weapon system."[351]

[347] Article 36 of 1977 Additional Protocol 1 to the Geneva Conventions.

[348] Ministry of Defence, 'UK Weapons Review' (*Ministry of Defence,* 11 March 2016) < https://assets.publishing.service.gov.uk/media/5a80bf5f40f0b62305b8cec5/20160308-UK_weapon_reviews.pdf > accessed 19 March 2019.

[349] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

[350] Ministry of Defence, 'UK Weapons Review' (*Ministry of Defence,* 11 March 2016) < https://assets.publishing.service.gov.uk/media/5a80bf5f40f0b62305b8cec5/20160308-UK_weapon_reviews.pdf > accessed 19 March 2019.

[351] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017

It is the view here that ‘subjective’ human decision-making is arbitrary and not something that can be quantified. It leads to ideas of soldiers going with their ‘gut-feelings’ and ignoring the factual situational data. MAMs, as a result of the underpinning technology, are highly likely to be more consistent and rational, evaluating all the data they have. Due to the sensors and their processing ability of highly complex data, they could run scenarios and outcomes far quicker than a human and provide an exact audit trail for a course of action, but this also presents a risk. The speakers at the 2014 Geneva argued that current technology, which presently includes heat sensors, visual sensors able to detect military uniforms or weapons, and sensors which can detect incoming fire, is not clever enough to make independent:

> “nuanced distinctions required by the principle of distinction, including distinguishing persons that are hors de combat from combatants, and civilians from those who are directly participating in hostilities. It is clear that the development of software that would be capable of carrying out such qualitative judgments is not possible with current technology.”[352]

Indeed, many speakers found the idea of when technology could make these judgements unimaginable, which is considered here as naive and reinforcing the argument of a future liability gap. The speakers had differing views on the suitability of IHL to control the development and usage of unconscious MAMs. Some speakers thought the existing law was sufficient, although saw value in having further guidance on testing, along with the need for legal reviews. Conversely, others stated that a definite ban on MAMs is necessary, or

[352] ICRC, ‘Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects’ (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017

establishment of a legal norm demanding, and defining, the term 'meaningful human control.'[353]

3.6 Summary

The chapter showed that there is a long- and well-established body of law for dealing with unconscious tools and entities that are operated by humans, or for organisations consisting of humans. These laws have served us well to date, however, the legal landscape will need to shift to accommodate AM development.

The chapter built on previous chapters around AM technology development and the resulting ethical considerations. Current legislation only focuses on AMs under human control and used akin to a tool, consequently, failing to address situations where an AM acts entirely on its own. It was shown that a psychological state can be attributed to a non-human legal entity such as a corporation. This approach considers the non-human legal entity as if it is competent of holding mental states, which, is argued here should now extend to include AMs, and would be the first step in recognising AMs rights.

Whilst AM consciousness is yet to be established, it was shown that the current law does not lay a path for its creation and recognition. It was highlighted that we should be mindful of the rights and duties we owe AMs, thus develop, or create, new legislation around this. It is

[353] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

further considered that ignoring a conscious entity is absurd, especially if we are responsible for its creation. It was highlighted that laws will need to be created or adjusted to embrace the development path of AMs, to acknowledge their advancement and to address the legal challenges and risks they will introduce. A first step to doing this would be to recognise them as a legal person, which legislation can flow from.

This chapter touched on civil law, with the purpose of setting the context of liability and demonstrating the current product related laws. The chapter moved its focus onto criminal liability, as criminal liability has the highest evidential bar and the core elements (actus reus and mens rea) have only been found in humans. The chapter highlighted the current gaps in criminal liability with regards to AMs, showing that the elements of criminal liability could indeed be met by AMs/MAMs in the future, thus liability should shift. Once AMs are conscious, they will be capable of setting their own goals and acting with free will, thus was argued that it would be unjust to hold the manufacturer or owner liable for an entity that they had no control over. As a result, it was asserted that it will be appropriate for AMs to be criminally responsible in their own right. In addition, it was pondered how the assessment of reasonableness would be quantified, and suggested that perhaps an AM on the Clapham omnibus style reasonableness test will need to be constructed.

Flowing from the view that AMs could meet the criminal liability requitements, Hallevy's three models for criminal liability were explored and shown to be a very attractive and flexible solution, which could be tailored to human-AM criminal liability as well as AM-AM liability. Hallevy's models were explored here as they were, at the time of writing, the only future thinking criminal liability framework that aimed to accommodate the development pathway to AM consciousness. It is the view here that Hallevy's models would only need to be slightly adapted for English law, with the inclusion of recklessness specifically for the first 2 of his models. His third model guards against future conscious AMs and allows for AMs to be assigned criminal responsibility, along with allowing them to have rights and duties.

Building on the non-military law, the chapter discussion moved to the military setting. It echoed chapter 2's discussion that IHL is born from the military action atrocities of human behaviour sanctioned by the State. As long as the BA adheres to IHL, then all actions, even human causalities, are lawful and no war crimes will be committed. A human soldier can commit a crime if they step outside the permitted action and/or in excessive of it, and shown here that this does not cover MAMs, creating a liability gap.

With regards to MAMs, there are significant gaps in the current legislation. Current legislation again does not regulate their development and views MAMs as being entirely under the control of a human during its operation and deployment. Further, there are key questions over how an MAM will be created and trained to understand, interpret and thus, adhere to complexities and subtleties of IHL. As a result, it is the opinion here that it could be considered unlawful and irresponsible to deploy MAMs into battle if accountability is not clarified, and the risks they pose not exposed and mitigated against.

The principle that 'accountability follows control'[354] is presently viewed with a human in mind, but could be applicable with MAMs, which could be held accountable for any breaches of IHL.

[354] Ministry of Defence, JSP 815. Element 5: Supervision, Contracting and Control Activities (JSP 815) Ministry of Defence < https://assets.publishing.service.gov.uk/media/66e18438dd4e6b59f0cb2500/JSP_815__Element_5_Supervision__contracting_and_control_activities_v1.2.pdf > accessed 4 September 2024.

However, this is unlikely to sit comfortably for many, as it could appear we are absolving ourselves of liability and any harm caused. When considering accountability for the use of MAMs, the chapter highlighted that the BA would need to consider their own responsibility in regard to decisions made, and their control MAMs, which is presently a significant gap. Indeed, the ethical and legal arguments around creating a conscious entity with the sole purpose of serving on the battlefield may ultimately prove too complex and therefore only MAMs without consciousness or legal personhood may be deployed into the battlefield. Consequently, the following chapters take this chapter's learning and identified gaps and considers the potential obligations and duties that could be afforded to MAMs, due to their legal personhood and consciousness. This could result in States being required to decide if conscious MAMs can ever be deployed into the battlefield and, if yes, adapt or draft new polices to include this.

Nevertheless, for there to be effective extrinsic control, it was discussed that the expectation of many scholars in this area, is that unconscious MAMs will stay under human control for many years due to the risks and political fallout of deploying a conscious entity without a human in the loop or 'meaningful human control'. MAMs will have advanced technology to aid human soldiers and protect their lives, so ultimately it will come down to a risk (both in terms of global power and legal), versus benefit analysis and the amount of risk and reputational damage the State is comfortable taking.

Through exploring the legal landscape and IHL, it has been shown that there is a significant liability gap arising with AMs and MAMs. AMs/MAMs will challenge our laws and possibly push them to breaking point, therefore we must be proactive and avoid a liability gap wherever possible. Thus, if we want to encourage innovation and technology development, then we be clear where liability lies. Further, the deployment of conscious autonomous military machines into conflict represents a value alignment problem (VAP) under IHL because it upends the assumptions of agency, consent, responsibility, and adherence to humanitarian principles that are fundamental to current legal and ethical frameworks. Addressing these issues requires not only a rethinking of legal accountability and control mechanisms but also a deeper philosophical examination of the moral status of conscious entities and their rights in the context of warfare; the true value alignment problem (TVAP). The TVAP will be discussed in the next chapter, as it shows how the development of machine consciousness gives rise to the TVAP argued by the author.

4. RA2: Exploration to understand if machine consciousness has been designed to work with IHL.

4.1 Introduction

This chapter delves deeper into the development of AI, looking at the technologies underpinning machine consciousness, what consciousness and machine consciousness is, and exploring if it has been designed to align to our human values[355] and work with IHL. It looks at the change from passive tool, to a self-thinking and directing entity, with its own objectives. it is argued here that designing machine consciousness to work with IHL is only as good as the human's initial intention, the values they hold, and the 'training' or 'nurturing' the AM receives, as, unlike a child, there is yet to be the safeguards implemented (e.g., a AMs version of social services). This chapter focuses on the responsibility of MAMs within the IHL landscape, which requires understanding of developing machine consciousness, the challenges, and human-robot relations for IHL. The chapter further explores the IHL considerations for conscious MAMs, the unique challenges posed by MAMs and what the author identifies as the true value alignment problem (TVAP).

[355] Anastasia Aldelina Lijadi , 'What are universally accepted human values that define 'a good life'? Historical perspective of value theory' (2019) WP-19-006 IIASA < https://pure.iiasa.ac.at/id/eprint/16049/1/WP-19-006.pdf > accessed 4 September 2024.

4.2 Overview of the Development of AI

Since the 1950s[356], there have been significant technological developments, which have led to autonomous machines being developed for use by the military, the medical profession, and more recently the development of autonomous vehicles (AVs). Semi-autonomous machines[357] have featured in manufacturing and mining industries for many years, although these machines are programmed to repeatedly perform a specific function under controlled conditions.[358] With the development of AMs, which think and act for themselves, (e.g., IBM's WatsonX[359]), and the move towards AVs, we are taking AMs out of the pre-defined, controlled conditions they are used to, and moving them into the chaos of the public arena, revealing novel legal regarding legal accountability and thus liability (as discussed in chapter 2 and 3, and ethical questions and concerns. Ethical challenges will be discussed in chapter 5 and 6.

4.3 Responsibility of AM/MAMs within the IHL Landscape

With the advent of machine Deep Learning (DL)[360], whereby machines act upon and evolve through stimuli and data[361], much like ourselves, questions have arisen over the legal status of an AM and liability for any harm they cause. The European Parliament's (EP) report into

[356] M. Haenlein and A. Kaplan, 'A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence' (2019) California Management Review, 61(4), 5-14. < https://doi.org/10.1177/0008125619864925 > accessed 4 September 2024.

357 From the literature read to date, the general consensus this this is Shared human and machine control.

358 Ugo Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts* (2013 edn, Springer).

[359] IBM, 'IBM Watson to Watsonx' (IBM, 2024) < https://www.ibm.com/watson?utm_content=SRCWW&p1=Search&p4=43700080376796564&p5=p&p9=58700008735428981&gad_source=1&gbraid=0AAAAAoS6_Rcu9_aAPTO1_7sLHdSUNLQ7U&gclid=CjwKCAjw68K4BhAuEiwAylp3krqhOUBDfHoNaTgaZ_QW1boUqB2rQndEWbVL392XtEKsVSwEUjzpdxoCfQYQAvD_BwE&gclsrc=aw.ds > accessed 4 September 2024.

[360] Explained later in the chapter.

[361] Michael Copland, 'What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?' (*Nvidia*, 29 July 2016) <https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/ > accessed on 24 April 2017.

Civil Law Rules on Robotics, highlighted that the EP are all too aware of the potential legal and ethical issues posed by the technological advances, stating that:

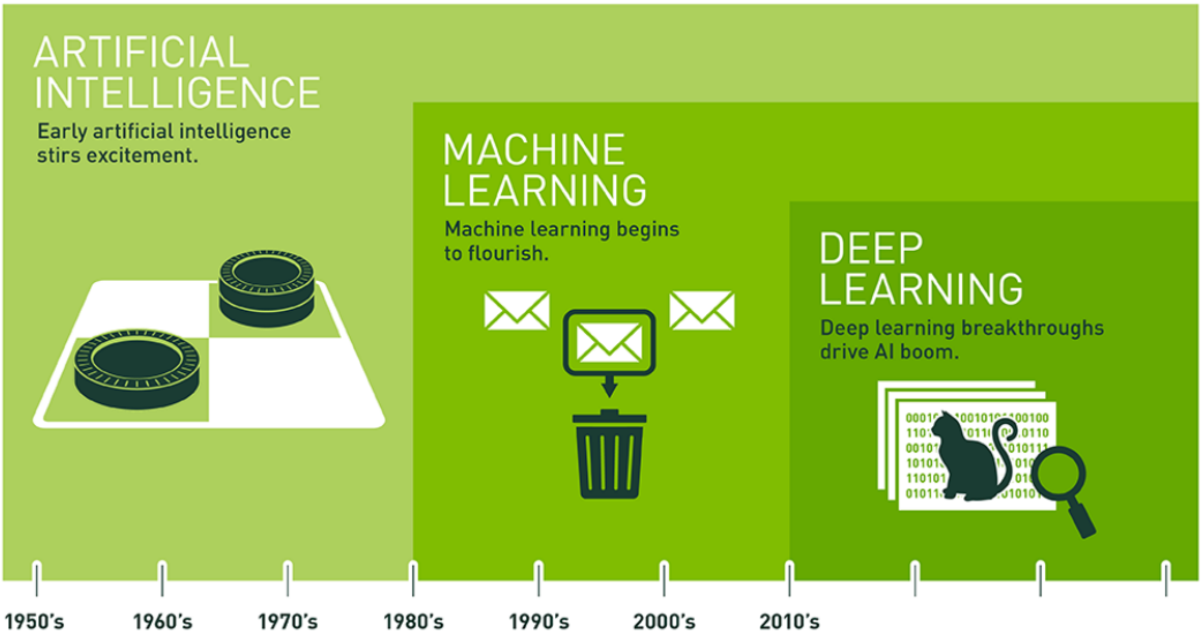
> “not only are today's robots able to perform activities which used to be typically and exclusively human, but the development of autonomous and cognitive features – e.g. the ability to learn from experience and take independent decisions – has made them more and more similar to agents that interact with their environment and are able to alter it significantly; whereas, in such a context, the legal responsibility arising from a robot’s harmful action becomes a crucial issue.”[362]

It is this advancement of machines making independent decisions through the dawn of DL, that makes this a problem of today and raises challenges for IHL, which will now be discussed. This is in stark contrast to the machines of yesterday, which took their commands from a human and acted as instructed or within set parameters.[363] The literature review drew focus to the development in the use of machines in wartime, which has significantly advanced from First and Second World Wars along with their capabilities, therefore, this chapter will not focus on machines from those times, but on the future developing technology of MAMs.

Since John McCarthy first introduced us to AI in 1956, the technology has significantly moved on. Today’s AMs go further into AI, expanding machine learning (ML) and into the newly developing world of DL, which will now be discussed. The diagram below provides an overview of the technological progression in this area.

[362] European Parliament, ‘*DRAFT REPORT with recommendations to the Commission on Civil Law Rules on Robotics*’ (2015/2103(INL)).

[363] Some semi-autonomous machines, for example vehicles, will be given an objective/goal, but the method of meeting that objective is their choice, e.g. a vehicle driving from A to B and deciding when to brake, speed and lane assist.


ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

Figure 1: AI Development.[364]

To reaffirm, this thesis does not look at the technical design standards and process (e.g. the British Defence Standards[365]), but rather explores the journey and arguments towards machine consciousness. Successively, machine consciousnesses and the value alignment problem (VAP) is considered as a result of the technological developments. In this context, the VAP relates to how we develop autonomous machines to behave and act in accordance with human norms and values, which can be nuanced.[366] The 'problem' is amplified and intensified when looking at the values underpinning IHL, as the potential consequences and loss could be devastating, which is discussed later in this chapter and chapter 6.

[364] Michael Copland, 'What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?' (*Nvidia*, 29 July 2016) <https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/ > accessed on 24 April 2017.

[365] Gov.UK, 'UK Defence Standardization' (Gov.UK, 30 January 2024) < https://www.gov.uk/guidance/uk-defence-standardization > accessed 4 September 2024.

[366] Brian Christian, *The Alignment Problem: How Can Artificial Intelligence Learn Human Values?* (September 2021, Atlantic Books)

4.4 Building the Brain: Developing Machine Consciousness

It is DL that is being used to develop autonomous, self-teaching machines that raise the challenges for IHL. DL is centred around algorithms called artificial neural networks, which originate from the complex structure and functions of our brain, and which is a stepping stone to machine consciousnesses. Data is fed into "a network's input nodes, which modify it and feed it to other nodes, and so on."[367] Nodes are the connection of artificial neurons together, which makes an architecture or processing structure.[368] This architecture forms a network, in which each input variable (called an input node) is connected to one or more output nodes.

As a reminder from the previous chapter, and to aid ease of reading, the author highlights again the relationship between DL, ML and AI, which are often described as a set of Russian dolls; DL is a subset of ML; ML is a subset of AI, that is an umbrella term for any smart computer program.[369] Natural Language Processing (NLP) is the practice of teaching machines to understand and interpret conversational inputs from humans.[370] Thus, all machine learning is AI, but not all AI is machine learning. Put simply, this means that ML is one method to achieve AI that on learning from data, but AI includes many different methods, which do not all include learning from data, for example, rule-based systems, which are pre-programmed

[367] Larry Hardesty, 'Making computers explain themselves' (*MIT News*, 27 October 2016) <http://news.mit.edu/2016/making-computers-explain-themselves-machine-learning-1028 > accessed on 24 April 2017.

[368] Robert Nisbet, Gary Miner, and Ken Yale, 'Chapter 7 - Basic Algorithms for Data Mining: A Brief Overview' (2018) Handbook of Statistical Analysis and Data Mining Applications (Second Edition), < https://doi.org/10.1016/B978-0-12-416632-5.00007-4 > accessed 4 September 2024.

[369] Kiltonsway, 'Artificial Intelligence (AI) vs. Machine Learning vs. Deep Learning' (Kiltonsway, 30 June 2021) < https://kiltonsway.mystrikingly.com/blog/artificial-intelligence-ai-vs-machine-learning-vs-deep-learning >accessed 13 June 2022.

[370] Ximena Bolaños 'Natural Language Processing and Machine Learning' (*Encora,* 29 September 2021) < https://www.encora.com/insights/natural-language-processing-and-machine-learning > accessed 4 September 2024.

with rules to make decisions, or symbolic AI, which use symbols and logic for knowledge and reasoning, such as chess algorithms.

The diagram below helps visualise the relationships that will be discussed and the progression towards machine consciousness, which is the focus of this thesis.

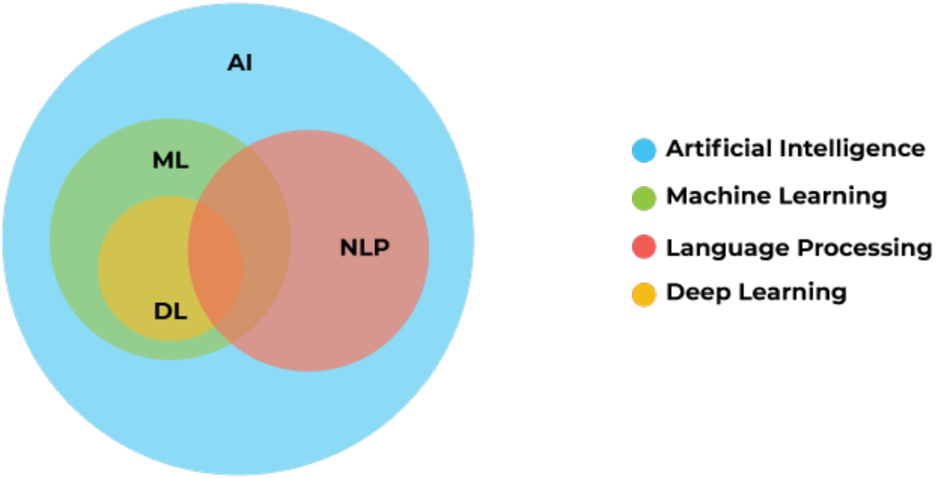


Figure 2: NLP and ML are part of AI and both subsets share techniques, algorithms, and knowledge.[371]

ML is a domain of computer science that focuses on the development of computer programs that can grow and teach themselves. Arthur Samuel,[372] an innovator in computer gaming, states that ML is the subclass of computer science that, "gives computers the ability to learn

[371] Ximena Bolaños 'Natural Language Processing and Machine Learning' (*Encora,* 29 September 2021) < https://www.encora.com/insights/natural-language-processing-and-machine-learning > accessed 4 September 2024.
[372] Arthur Lee Samuel (December 5, 1901 – July 29, 1990) was an leading American expert in computer gaming and artificial intelligence.

without being explicitly programmed."[373] ML enables developers to build algorithms that find patterns in existing data and automatically improve themselves without instructions from the developer or any other human. It relies exclusively on the data, therefore the more data available, the more efficient ML, yet it is important to ensure the quality and validity of the data, as the learning and thus feedback, could quickly result in 'bad' decisions and actions being taken. This is of particular concern when considering IHL related decisions, and the consequences of not making decisions that are aligned to IHL values, as expected. This is simplified in Fig 3 below, which shows the learning loop for how the values of IHL will be learnt.

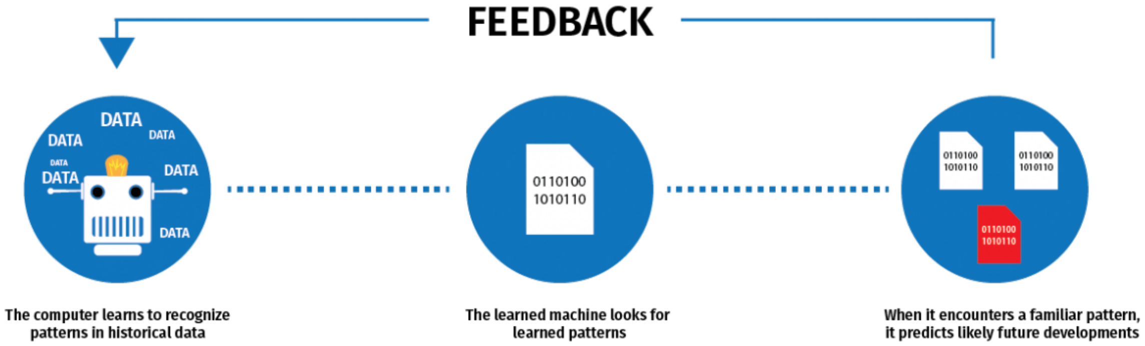

FEEDBACK
DATA
DATA
DATA
DATA
DATA
DATA
DATA
0110100
1010110
0110100
1010110
0110100
1010110
0110100
1010110
The computer learns to recognize patterns in historical data
The learned machine looks for learned patterns
When it encounters a familiar pattern, it predicts likely future developments

Figure 3: Feedback Learning Loop.[374]

DL is centred around algorithms called artificial neural networks (ANNs), which originate from the complex structure and functions of our brain. Drawing on the capability of ANNs, it is argued here that AMs will reach a degree of consciousness within our lifetime. Whilst decision

[373] M Awad and R Khanna, 'Machine Learning. In: Efficient Learning Machines' (2015) Apress, Berkeley, CA. < https://doi.org/10.1007/978-1-4302-5990-9_1 > accessed 10 November 2017.
[374] Synectics, 'Evolution of Machine Learning' (Synetices, 2018) < http://www.smdi.com/evolution-machine-learning > accessed 3 March 2020.

making is the focus of chapter 5, the technology and process development, thus how they think, is heavily linked to consciousness. This development, along with the level of technological ambition, has led to the potential challenges that will now be discussed. It also serves to highlight the potential challenges in training MAMs at scale, and to a IHL approved universal standard.

4.5 The Challenges of Learning and Achieving Goals

4.5.1 Machine Learning Challenges for IHL

To appreciate the challenges IHL may face, it is necessary to have a comprehensive understanding of the technology, the trajectory and the associated risks, from the significant advancement of non-military technology. To help understand the underpinning technology and learning complexities, a good starting point is ML. ML involves learning from data inputs, then evaluating and optimising it to create actions or outputs. It is commonly used in data analytics as a way to develop algorithms for making data predictions and is associated with probability, statistics, and linear algebra. It is generally classified into three areas dependent upon the type of the learning signal or feedback accessible to a learning system.[375]

1. Supervised learning: The machine is shown inputs and the desired outputs, with the goal of learning a general rule to map the inputs to the outputs.

[375] Synectics, 'Evolution of Machine Learning' (Synetices, 2018) < http://www.smdi.com/evolution-machine-learning > accessed 3 March 2020.

2. Unsupervised learning: The machine is shown inputs but without the desired outputs, and the goal is to find structure in the inputs.
3. Reinforcement learning: The machine works with a dynamic environment to perform a specific goal without guidance or human help.

ML maximises the machines' ability to learn from relationships hidden in the data and can be exploited further through the development of intelligent and effective machine learning algorithms. To stress, these are unconscious AMs used in a non-military setting, although, relevant here as the technology development pathway is the foundations of MAMs.

Whilst it may appear new, ML has been around for some time, yet has now come into production due to lower technology costs and being more accessible e.g., ChatGPT. It has progressed to decipher real life problems, and automate processes, used in various sectors such as healthcare, banking, retail, etc. The software, application or solution, developed via ML can adapt to changing requirements through learning from its dynamic environment. Adaption for unconscious-MAMs is acceptable when a human is in control, holds the required values, and decides the action to take, but for conscious MAMs outside of human control, this is a significant risk to IHL, if the MAM does not understand the nuances of IHL and the values underpinning it when making decisions. This situation can lead to a misalignment with human values, including those underpinning IHL, and is the basis of the VAP. The VAP focuses on designing AI so that its "goals and behaviours can be assured to align with human values throughout their operation"[376] (discussed further in chapter 5). Nevertheless, the lessons learnt from ML can be scaled up and shared across multiple non-military applications,

[376] Gabriel, I, 'Artificial Intelligence, Values, and Alignment' (2020) Minds & Machines **30**, 411–437 < https://doi.org/10.1007/s11023-020-09539-2 > accessed 4 September 2014.

reducing the risks and refining the learning. Indeed, ML inherently reflects upon many variables that impact the results or observations.

4.5.1.1 Natural Language Processing (NLP) and IHL

Natural language processing (NLP) facilitates the machine to execute tasks and automate manual processes from human input that could be in text or audio form, via 5 stages.[377] This process is good for where an AM's 'thinking' is within a controlled environment, and where mistakes are unlikely to be deadly, thus, this may not be suitable for MAMs within the military context. Indeed, learning the rules in a military setting, under the governance of IHL, means misunderstanding and misinterpretation must be zero, and, as argued in chapter 5, MAMs are unlikely to be afforded learning and development time, which may result in a misalignment with human values.

4.5.1.2 Deep Learning (DL) and IHL

DL is a part of a wider group of DL methods that are also known as 'deep structured learning'[378], hierarchical learning, or Deep ML. ML programs modify their response according to the data they are exposed to experience. Artificial neural network (ANN), which is a machine learning algorithm, has been renamed as DL. To note, DL is a subset of ML. However, people often use the term to refer to deep ANNs, or occasionally for deep reinforcement learning.

[377] Synectics, 'Evolution of Machine Learning' (Synetices, 2018) < http://www.smdi.com/evolution-machine-learning > accessed 3 March 2020.

[378] Synectics, 'Evolution of Machine Learning' (Synetices, 2018) < http://www.smdi.com/evolution-machine-learning > accessed 3 March 2020.

ANNs are a category of machines, which are a result of analysing the structure and function of the human brain. They are developed from simplistic processing nodes shaped into a network. Essentially, they are pattern recognition systems and lend their usefulness to tasks that can be described regarding pattern recognition.[379] They are taught via feeding them datasets of known outputs.

Despite these advancements, machines still make errors. However, errors can be minimised by constructing a framework that multiplies inputs so to make guesses as to the inputs' type. Various outputs (guesses) are the result of the inputs and the algorithm. Normally, the early guesses are very wrong (the software engineer/programmer knows the correct answer), although this is part of the learning process. However, if there are baseline markers relating to the input, then the incorrect guesses can be measured by comparing them with the baseline. This comparison can be used to modify the algorithm. This is exactly what neural networks achieve. Neural networks continue to measure errors and modify the parameters until they get to a point where a less error result cannot be achieved. Put simply, they are an optimisation algorithm. Tweaked correctly, they will minimise error by continually guessing, however, error in relation to IHL could be devastating and contrary to our values (the VAP), which again, raises the question here, if MAMs will be afforded 'learning time' to learn. It is the view here that training time is necessary to address both the VAP and the TVAP by demonstrating that we accept the training burden placed on us and the magnitude of the task, plus acknowledge that there is a need for shared learning of both human and MAM values. This learning includes time to understand how to interpret, comprehend and

[379] Synectics, 'Evolution of Machine Learning' (Synetices, 2018) < http://www.smdi.com/evolution-machine-learning > accessed 3 March 2020.

formulate an appropriate response that aligns with human, including IHL, values. Further, the author is unsure what, if any, level of mistake would be deemed acceptable.

Deep ANNs are a group of algorithms, which are known for their accuracy in solving many complex problems, for example, image recognition, sound recognition, and natural language processing, and would therefore be highly beneficial to military machines. Indeed, this would enable them to process multiple sources of data and suggest military operational options/scenarios. It is all these features of an ANN that can have a considerable effect on the performance of the model and its decision making, along with the quality and type of data, it is fed and the expertise of the programmer. Thus, if the data quality or quantity is poor, then the decision-making will be poor, resulting in inappropriate and even incorrect actions and outcomes.

As highlighted, ANNs are very powerful, however, they are also very complex and deemed 'black box algorithms' due to being very hard to understand and explain. Further, Bruiger states,

> "to presume to control the evolution of such a system (a black box) by setting its "initial conditions" is problematic if not paradoxical, since what black box contains cannot be presumed to be a deterministic system. It can be known only by its observed outputs. It can be controlled only by containing what we think are its inputs and outputs – which is how we generally deal with physical systems, other creatures, and people."[380]

This should be borne in mind when looking to use them to solve problems, as unpicking why an AM did, or did not, do something will prove impossible, which introduces substantial risk.

[380] Dan Bruiger, 'The Value Alignment Problem' (2021) PhilPapers < https://philpapers.org/versions/BRUTVA > accessed 4 September 2024.

This will have significant implications for IHL and meaningful control, as IHL creates "obligations for human combatants in the use of weapons to carry out attacks, and it is combatants who are both responsible for respecting these rules, and who will be held accountable for any violations."[381] Indeed, it is argued here that responsibility and meaningful control will be near impossible if a MAM was acting on its own decision or interpretation. However, human soldiers, just like all other humans, also share 'black box' characteristics, in that we cannot see exactly why a human did or did not do something. Thus, shared ideology and IHL value alignment in human soldiers is just as important as MAMs, although one hopes this would have factored into the human soldier's enlisting decision making.

ANNs and the more complex DL methods are the best AI tools for solving extremely complex problems, however, they do still face challenges. These challenges include a propensity to forget patterns, the computational power they require, transferring learning from one context to another, and, as just mentioned, probably the most concerning with regards to IHL, that neural networks operate like black boxes, so not suitable where things/actions need to be explained. Consequently, to solve the challenges requires looking beyond deep neural networks and to potential solutions such as cogency maximization, adaptive resonance theory and neuro-fuzzy systems.

Adaptive resonance theory (ART), a type of neural network technique, is "a cognitive and neural theory of how the brain autonomously learns to categorize, recognize, and predict

[381] Neil Davison, 'A legal perspective: Autonomous weapon systems under international humanitarian law' (2018) UNODA Occasional Papers No. 30 < https://www.icrc.org/sites/default/files/document/file_list/autonomous_weapon_systems_under_international_humanitarian_law.pdf > accessed 4 September 2024.

objects and events in a changing world."[382] It tackles the issue of stability versus plasticity (e.g., developing new learning without losing existing knowledge) and is also called incremental or online learning. The major benefit of ART is that it does not need to retrain its model, which means it does not lose data, which is often a problem with retraining. Not losing data could also introduce problems, as humans change how they perceive something or change their views, which may not prove so easy in an AM or MAM, for example, viewing a State as friendly after a period of hostility.

It is the opinion here that being able to change a perception on something, helps humans stay aligned to IHL and the underpinning values. Further, it is seen here as a significant risk if MAMs acted on out-of-date data, which does not align to IHL values and thus exacerbates the VAP. This looks to impart an obligation on the users and/or trainers of MAMs to ensure their data is current and aligned to IHL values.

As an aside, the author asserts that because of this 'change a perception' human attribute, it could aid the addressing of the TVAP for MAMs with regards to IHL, and ensure their 'life' is recognised and protected.

IBM has integrated deep learning into its 'cognitive computing' systems, such as Watson.[383] Cognitive computing, IBM's term to describe machines that do not require explicit

[382] Stephen Grossberg, 'Adaptive Resonance Theory: How a brain learns to consciously attend, learn, and recognize a changing world' (2013) Neural Networks, Volume 37 < https://doi.org/10.1016/j.neunet.2012.09.017 > accessed 4 September 2024.

[383] IBM, 'IBM Watson to Watsonx' (IBM, 2024) < https://www.ibm.com/watson?utm_content=SRCWW&p1=Search&p4=43700080376796564&p5=p&p9=58700008735428981&gad_source=1&gbraid=0AAAAAoS6_Rcu9_aAPTO1_7sLHdSUNLQ7U&gclid=CjwKCAjw68K4BhAuEiwAylp3krqhOUBDfHoNaTgaZ_QW1boUqB2rQndEWbVL392XtEKsVSwEUjzpdxoCfQYQAvD_BwE&gclsrc=aw.ds > accessed 4 September 2024.

programming, builds on ML and NLP and aims to be capable of problem solving without human assistance or guidance.[384] Training is via setting performance measures, which improve over time and via data feedback, for example, a banking fraud detection system; After the banking transaction has been processed, it is known if it is fraudulent or not, so then this can be fed back into the system for it to continue to learn from. Thus, it is vital to think carefully about the choice of performance measure and choose one that would suit the needs and risk appetite of the organisation for their MAM, which should be higher than the risk tolerance of IHL, otherwise it could be considered that the organisation went against or undermined the IHL principes.

4.6 How the Goals Are Achieved

4.6.1 Machine Learning Algorithms

The above sections cover how to set goals for ML and the associated challenges. However, they do not show how to actually achieve the goal, which are typically accomplished via the most popular ML algorithms, which are:[385]

- Supervised Learning - The algorithm is provided with training data that contains the 'correct answer' for each example.[386] This is teaching method when the answer is

[384] Synectics, 'Evolution of Machine Learning' (Synetices, 2018) < http://www.smdi.com/evolution-machine-learning > accessed 3 March 2020.

[385] IBM Website, 'What is machine learning?' (IBM.COM) <https://www.ibm.com/topics/machine-learning?mhsrc=ibmsearch_a&mhq=what%20is%20machine%20learning > accessed 19 October 2020.

[386] Jafar Alzubi, Anand Nayyar and Akshi Kumar, 'Machine Learning from Theory to Algorithms: An Overview' (2018) J. Phys.: Conf. Ser. 1142 012012 <https://iopscience.iop.org/article/10.1088/1742-6596/1142/1/012012/meta?gclid=CjwKCAjwsKqoBhBPEiwALrrqiP3vMzDx9JGz1TANRsXtG34CXZmnRZDEY75gv0B6nf6tXjitLH8tPBoCvS4QAvD_BwE> accessed 3 March 2020.

binary, yet the 'correct answer' with regards to IHL may not always be as such. IHL answers and the resulting action to take, is often nuanced and subjective.

- Unsupervised Learning - The algorithm examines the training data for structure. For example, identifying the data examples that are like each other, then grouping them together.[387] However, this could cause unintended problems for IHL, resulting in biases creeping in. For example, problems with distinguishing between non-military and military combatants. There would also be a significant cost both in terms of money and effort, to create sufficient training data.

The above algorithms highlight how MAMs will learn and emphasises challenges of overlaying IHL. It will be impossible to teach MAMs every situation and scenario they could face, thus their interpretation may differ from a human soldier. Further, they may not perceive or see the value of the desired outcome in the same way a human soldier or the overall State. Again, this is where MAM value alignment could diverge and be at odds with IHL. It is the view here that potential diverge is likely to be due to their interpretation of a situation/data and their 'lived' experience, which, due to being designed and trained for a specific purpose, will only have a military context, thus not balance with a civilian/non-military context.

[387] Jafar Alzubi, Anand Nayyar and Akshi Kumar, 'Machine Learning from Theory to Algorithms: An Overview' (2018) J. Phys.: Conf. Ser. 1142 012012 <https://iopscience.iop.org/article/10.1088/1742-6596/1142/1/012012/meta?gclid=CjwKCAjwsKqoBhBPEiwALrrqiP3vMzDx9JGz1TANRsXtG34CXZmnRZDEY75gv0B6nf6tXjitLH8tPBoCvS4QAvD_BwE> accessed 3 March 2020.

4.6.2 Improvement of Machine Learning

One way to further improve the task of ML is to look at the types of problems it can solve. Some of the most frequent ones are:

- “Regression - A supervised learning problem where the answer to be learned is a continuous value. For instance, the algorithm could be fed with a record of house sales with their price, and it learns how to set prices for houses.”[388]
- “Classification - A supervised learning problem where the answer to be learned is one of finitely many possible values.”[389] In the banking fraud example, the algorithm must learn how to determine between 'fraud' and 'honest' to respond in the correct way. This is an example of a binary problem as it has only two possible values.
- “Segmentation - An unsupervised learning problem where the structure to be learned is a set of clusters of similar examples. For instance, market segmentation aims at grouping customers in clusters of people with similar buying behavior.”[390]
- “Network analysis - An unsupervised learning problem where the structure to be learned is information about the importance and the role of nodes in the network. For instance, the page rank algorithm analyzes the network made of web pages and their hyperlinks and finds what are the most important pages. This is used in web search

[388] IBM Website, 'What is machine learning?' (IBM.COM) <https://www.ibm.com/topics/machine-learning?mhsrc=ibmsearch_a&mhq=what%20is%20machine%20learning > accessed 19 October 2020.

[389] IBM Website, 'What is machine learning?' (IBM.COM) <https://www.ibm.com/topics/machine-learning?mhsrc=ibmsearch_a&mhq=what%20is%20machine%20learning > accessed 19 October 2020.

[390]IBM Website, 'What is machine learning?' (IBM.COM) <https://www.ibm.com/topics/machine-learning?mhsrc=ibmsearch_a&mhq=what%20is%20machine%20learning > accessed 19 October 2020.

> engines like Google. Other network analysis problems include social network analysis."[391]

ML can help with more problems than those just listed above, however, the above are typical problems. It should be noted that these problems are about data structure and classification, with a human overseeing the resulting action. These problems are beneficial for ML to help with where the risk of harm is low, e.g., predicting house prices. At its most challenging, MAMs will need to process data and take action in a chaotic environment, which may not afford the luxury of human review time. However, oversimplifying the military environment could lead to generalisations, misalignment with our values, resulting in incorrect ILHL applications and actions taken. As a result, it is asserted here that the need to develop and demonstrate the correct application of IHL and value alignment through controlled training environments, is crucial.

4.6.3 Machine Learning Workflow

The problem with the definitions above, is that developing a ML algorithm alone is not suitable for creating a system that learns. There is a disparity between a machine learning algorithm and a learning system.[392]

The diagram below shows the flow of machine learning:

[391] IBM Website, 'What is machine learning?' (IBM.COM) <https://www.ibm.com/topics/machine-learning?mhsrc=ibmsearch_a&mhq=what%20is%20machine%20learning > accessed 19 October 2020.

[392] IBM Website, 'What is machine learning?' (IBM.COM) <https://www.ibm.com/topics/machine-learning?mhsrc=ibmsearch_a&mhq=what%20is%20machine%20learning > accessed 19 October 2020.

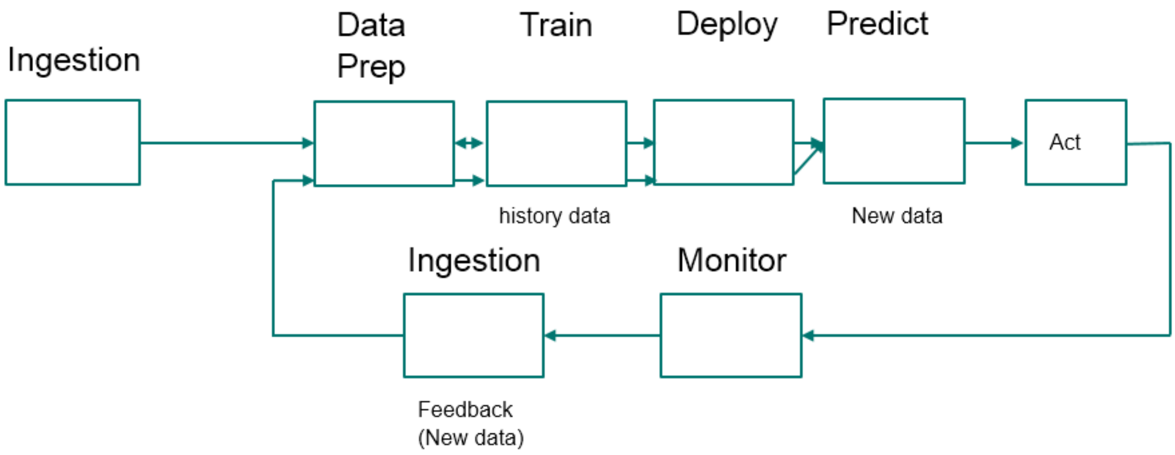


Figure 4: Flow of Machine Learning.[393]

The ‘Train’ step is where a ML algorithm is used. The output (a trained model) is then utilised by the 'Predict' stage of the workflow. It is the quality of the predictions from the ‘Predict’ stage that differentiates between a good and a bad machine algorithm. Thus, “the purpose of machine learning is to learn from training data in order to make as good as possible predictions on new, unseen, data.”[394] A big ML problem, and resultingly a significant risk, is to be able to develop a model that can lead to a good prediction on unforeseen data, thus:

> “how can we evaluate the quality of a model without looking at the data on which we will make predictions?...The general idea is that we assume that unforeseen data is similar to the data we can see. If a model is good on the data we can see, then it should be good for unforeseen data.”[395]

This allows for decisions to be made based on assumptions. To stress, AM development is a fast growing and extremely powerful area, which will have a significant impact on the day-to-

[393] IBM Website, 'What is machine learning?' (IBM.COM) <https://www.ibm.com/topics/machine-learning?mhsrc=ibmsearch_a&mhq=what%20is%20machine%20learning > accessed 19 October 2020.
[394]IBM Website, 'What is machine learning?' (IBM.COM) <https://www.ibm.com/topics/machine-learning?mhsrc=ibmsearch_a&mhq=what%20is%20machine%20learning > accessed 19 October 2020.
[395] IBM Website, 'What is machine learning?' (IBM.COM) <https://www.ibm.com/topics/machine-learning?mhsrc=ibmsearch_a&mhq=what%20is%20machine%20learning > accessed 19 October 2020.

day workings of modern society. The fact they learn how to learn means we need to be careful of the data we expose them to (and argued here as our nurturing requirement), as we will lose control of how the data is used and interpreted via DL. Further, how well received MAMs operating on assumptions in theatre will be, along with how this aligns to our values and IHL, specifically the principles of distinction, proportionality, and necessity, is yet to be determined. It is the opinion here that assumptions risk undermining IHL and our values. This in turn could weaken the UK's reputation of being an ardent supporter and enforcer of IHL. Thus, it is the opinion here that uncertainty could be reduced through rigorous testing for both humans and MAMs in a virtual environment that simulates the battlefield.

4.7 Machine Consciousness and The Challenges

4.7.1 Understanding Consciousness

The previous sections have focused on how goals are set and subsequently achieved, thus how 'thinking' machines created, however, the discussion has shown that these are without personhood and recognised autonomy in mind. It is this trajectory, driven by technologist and industry, that denotes that AMs are transforming before the ink on their history can dry. As a result, projections on the future of AMs, including their capabilities, are becoming outdated just as fast. This rapid development will have law makers and law enforcers on the backfoot and playing catch up, specifically in recognising and acknowledging the need to extend IHL protection to MAMs, and for their benefit. This is argued here as the true value alignment problem (TVAP).

Consciousness is vital to human agency and thus will be a key requirement of autonomy within AMs and MAMs. Designing and developing machine consciousness will be one major feat, but recognising the implications and consequences is an assertion here as the biggest challenge and where the TVAP arises. Indeed, establishing consciousness initiates the argument for recognising personhood. This subsequently leads to realising agency and free will, which constitutes autonomy. This ethical facet of machine consciousness is discussed next, whilst personhood, autonomy, and agency are discussed in chapter 5.

4.7.2 The Theory of Value Alignment

This chapter has introduced the reader to IHL and value alignment, however, this will now be discussed in more detail in the following section.

When considering the VAP in relation to AMs aligning to human values, Bruiger stresses:

> "To follow or obey a command is a different action for an AI agent than it is for an AI tool. An agent decides for itself how to interpret the input (in light of its own goals) and whether and how to respond to the command…A machine "obeys" a command automatically, with no intervening will and no goals of its own. An agent may or may not embrace the programmer's goal as its own, weighed against the backdrop of the agent's own priorities."[396]

Further, Bruiger views "the dilemma of imparting stable goals is that a self-modifying AI may modify what the programmer has initially specified as its goal."[397] He goes on to state, "if we suppose that it "naturally" modifies its goal through "reasoned understanding", then we must

[396] Dan Bruiger, 'The Value Alignment Problem' (2021) PhilPapers < https://philpapers.org/versions/BRUTVA > accessed 4 September 2024.

[397] Dan Bruiger, 'The Value Alignment Problem' (2021) PhilPapers < https://philpapers.org/versions/BRUTVA > accessed 4 September 2024.

presume an agent with its own purposes. In that case, the human goals concerned must be negotiated with the AI, as they would be with other human or animal agents. Moreover, an AI that is not an agent cannot "care" about anything, including its own effectiveness."[398] This is key, as we will want AMs to care about our values and to care about upholding IHL. To do this, it is stated here that we must negotiate the benefits of IHL with MAMs and help them see for themselves the value. Indeed, it is the view here that value alignment for unconsciousness MAMs may be possible, yet for consciousness MAMs, this should not be considered a given and something we can indoctrinate. MAMs will need to recognise the benefit to caring about our values and IHL, which is held here will not automatically be the case when they develop consciousness; they will need to see for themselves why they should care. This is discussed further in chapter 5.

Sheridan[399], Hendriks[400] and Darling[401] emphasise the link between AM consciousness, increase of social robots (e.g., Sophia) and human-robot social interaction. This interaction can lead the human to having an emotional connection, especially due to our trait of

[398] Dan Bruiger, 'The Value Alignment Problem' (2021) PhilPapers < https://philpapers.org/versions/BRUTVA > accessed 4 September 2024.

[399] T B Sheridan, T. B, 'Human-robot interaction: status and challenges' [2016] Hum. Factors 58, 525–532. doi: 10.1177/0018720816644364.

[400] B Hendriks, B Meerbeek, S Boess, S Pauws, and M Sonneveld, 'Robot vacuum cleaner personality and behavior' [2011] Int. J. Soc. Robots 3, 187–195. doi: 10.1007/s12369-010-0084-5.

[401] K Darling, 'Extending legal protection to social robots: the effects of anthropomorphism, empathy, and violent behavior towards robotic objects' (2012) We Robot Conference 2012, April 23, 2012, < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2044797 > accessed 10 November 2017.

anthropomorphising.[402] When looking at consciousness, philosophers Hegal[403] and Kant[404] link morality to ideas of consciousness, personhood, free will and rationality, which underpin the TVAP as argued by the author. These ideas are heavily linked to ethical theories and thus explored in detail in chapter 5. Exploring further links to consciousnesses, such as Lockes'[405] assertion that intelligence is to combine with consciousness and , it becomes quickly apparent to the author that there is no clear definition or measure of intelligence. Nilsson[406] and Stone et al[407] have looked at this from an AI perspective, although stop short of defining a truly universal and all-encompassing AM intelligence term.

The infamous Turing test, which focus on assessing a machine's ability to display human like responses and intelligence,[408] is criticised by Searle[409], with Signorelli and Arsiwalla[410] proposing that it is better to create a test based around moral dilemmas rather than simple everyday questions. For morality, and moral responsibility (discussed in chapter 5), there is a requirement for high-level cognition, which allows for self-reflection and understanding of

[402] M Scheutz, "The inherent dangers of unidirectional emotional bonds between humans and social robots," [2011] Robot Ethics, The Ethical and Social Implications of Robotics, eds P. Lin, K. Abney, and G. A. Bekey (MIT Press),
K Darling, 'Extending legal protection to social robots: the effects of anthropomorphism, empathy, and violent behavior towards robotic objects' (2012) We Robot Conference 2012, April 23, 2012, < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2044797 > accessed 10 November 2017.
K Darling, 'Extending legal protection to social robots: the effects of anthropomorphism, empathy, and violent behavior towards robotic objects' (2012) We Robot Conference 2012, April 23, 2012, < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2044797 > accessed 10 November 2017,
D J Gunkel, 'Robot Rights' [2018] MIT Press. doi: 10.7551/mitpress/11444.001.0001,
Fisher, 1991, Disambiguating anthropomorphism: An interdisciplinary review. J.A. Fisher Perspectives in Ethology, 9 (1991), pp. 49-85.

[403] G W F Hegel, 'Philosophy of Right' (2001) Transition (Vol. 1), Kitchener, Batoche Books Limited.

[404] I Kant, *Fundamental Principles of the Metaphysic of Morals*, (1785) 1949th ed. New York, NY: L. A. Press, Ed.

[405] John Locke was an English philosopher and physician (29 August 1632 – 28 October 1704).

[406] N J Nilsson, 'The Quest for Artificial Intelligence' (2009) Cambridge University Press < https://ai.stanford.edu/~nilsson/QAI/qai.pdf > accessed 10 November 2017.

[407] P Stone, R Brooks, E Brynjolfsson, R Calo, O Etzioni, G Hager, et al, 'Artificial Intelligence and Life in 2030' (2016) One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel. Stanford < http://ai100.stanford.edu/2016-report > accessed 10 November 2017.

[408] Stanford Encyclopedia of Philosophy, 'Turing Test' (Stanford Encyclopedia of Philosophy, 4 October 2021) < https://plato.stanford.edu/entries/turing-test/ > accessed 4 September 2024.

[409] J R Searle, *Minds, brains, and programs* (1980) Behav. Brain Sci. 3.

[410] C M Signorelli, and X D Arsiwalla, 'Moral Dilemmas for Artificial Intelligence: a position paper on an application of Compositional Quantum Cognition' [2018] Quantum Interaction. QI 2018. Lecture Notes in Computer Science (Nice).

mistakes.[411] Moral principles and values affect how we make decisions and lead our lives, and thus the decisions, value and compliance with IHL. When examining the requirement of consciousness for human intelligence, Barron and Klein[412] highlight that 'subjective experience' is created from emotional and rational intelligence. Baars[413], Tononi and Koch[414] assert that humans are required to be conscious in order to make complex rational decisions, and have the required intention to do something, thus vegetative patients and minimally conscious patients do not meet these requirements nor show any intention to perform minimal tasks[415], despite, at times, showing minimal signs of consciousness, such as a response to stimuli, such as touch. [416] These responses, no matter how small, show the brain is reacting and raises questions over relax responses verses an element of communication. Whilst of interest, this neurological response area is not debated further.

Varela[417], Kauffman and Varela[418], Kauffman[419] argue that there are no less than two core processes inherent in consciousness: 1) awareness, and 2) self-reference, which align to requirements for autonomy (discussed in chapter 5) and consequently lead to accountability under IHL. Varela and Goguen proclaim that consciousness materialises from the entirety of

[411] W J Gehring, B Goss, M G H Coles, D E Meyer, and E Donchin, 'A neural system for error detection and compensation' [1993] Psychol. Sci. 385–390. doi: 10.1111/j.1467-9280.1993.tb00586.x,
J D Smith 'The study of animal metacognition' [2009] Trends Cognit. Sci. 13, 389–396. doi: 10.1016/j.tics.2009.06.009,
S M Fleming, R S Weil, Z Nagy, R J Dolan, and G Rees, 'Relating introspective accuracy to individual differences in brain structure' [2012] Science 329, 1541–1544. doi: 10.1126/science.1191883.

[412] A B Barron and C Klein, 'What insects can tell us about the origins of consciousness' [2016] Proc. Natl. Acad. Sci. U.S.A. 113, 4900–4908. doi: 10.1073/pnas.1520084113.

[413] B J Baars, 'Global workspace theory of consciousness: toward a cognitive neuroscience of human experience' [2005] Prog. Brain Res. 150, 45–53. doi: 10.1016/S0079-6123(05)50004-9.

[414] G Tononi and C Koch, 'The neural correlates of consciousness: an update' [2008] Ann. N. Y. Acad. Sci. 1124, 239–261. doi: 10.1196/annals.1440.004.

[415] O Gosseries, H Di, S Laureys,and M Boly, 'Measuring consciousness in severely damaged brains' [2014] Annu. Rev. Neurosci. 37, 457–478. doi: 10.1146/annurev-neuro-062012-170339.

[416] A M Owen, M R Coleman, M Boly, M H Davis, S Laureys, and J D Pickard, (2006) 'Detecting awareness in the vegetative state' [2006] Science 313:1402. doi: 10.1126/science.1130197.

[417] F J Varela, 'A calculus for self-reference' [1975] Int. J. Gen. Syst. 2, 5–24.

[418] L H Kauffman, and F J Varela, 'Form dynamics' [1980] J. Soc. Biol. Syst.3, 171–206. doi: 10.1016/0140-1750(80)90008-1.

[419] L H Kauffman, 'Self-reference and recursive forms' [1987] J. Soc. Biol. Syst. 10, 53–72. doi: 10.1016/0140-1750(87)90034-0.

the processes.[420] Deshmukh states that consciousness and awareness are not the same. As a result, Deshmukh views that consciousness is a "cognitive and dualistic process"[421], whilst awareness is "nondual, spontaneous, and nonlocal."[422] With regards to self-reference, it is the "action of the Self of looking back at itself"[423], which Visan argues this is "the only self-reference that can exists, for the reason that the Self is the only real entity that can refer back to itself.[424] These theories and debates serve to highlight the challenges of defining consciousnesses in humans, therefore stress the enormity of the challenges in recognising it in AMs/MAMs.

Locke's idea of self-consciousness combined with basic intelligence, set a precedent for today's thinking on personhood and established these as the key features. Locke's idea that personal identity (or the 'self') is founded on consciousness and not a construct of either the soul or the body, is again transferrable to AMs, in that he ignores the major barrier of the physical body and focuses on consciousness alone, nevertheless, it ignores the fundamental flaw of human biology and fragility, which makes us arguable more vulnerable than MAMs. Harris echoes Locke's thinking, yet focuses the argument on the importance of allowing an individual to value their own existence, saying:

> "The important feature of this account of what it takes to be a person, namely that a person is a creature capable of valuing its own existence, is that it also makes plausible an explanation of the nature of the wrong done to such a being when it is deprived of existence."[425]

[420] F J Varela, and J A Goguen, 'The arithmetic of closure' [1978] Cybernet. 8, 291–324. doi: 10.1080/01969727808927587.

[421] VD Deshmukh, 'Consciousness, Awareness, and Presence: A Neurobiological Perspective' (2022) Int J Yoga 2022 May-Aug < https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9623886/ > accessed 4 September 2024.

[422] VD Deshmukh, 'Consciousness, Awareness, and Presence: A Neurobiological Perspective' (2022) Int J Yoga 2022 May-Aug < https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9623886/ > accessed 4 September 2024.

[423] Cosmin Vișan, 'The Self-Referential Aspect of Consciousness' (2017) Journal of Consciousness Exploration & Research Vol 8 Iss 11 < https://philarchive.org/archive/VISTSA-2 > accessed 4 September 2024.

[424] Cosmin Vișan, 'The Self-Referential Aspect of Consciousness' (2017) Journal of Consciousness Exploration & Research Vol 8 Iss 11 < https://philarchive.org/archive/VISTSA-2 > accessed 4 September 2024.

[425] J. Harris, 'Wonderwoman and Superman,' (1992) Oxford.

Locke was unlikely to have the concept of machine consciousness in mind when writing his essay, yet his seedling ideas, along with their applicability and reach, are explored in this thesis. Indeed, Locke's ideas are used in this thesis to argue that AMs will value their existence and appreciate the contribution they make to their world, e.g., advancing their own AM development and knowledge, which could stretch the remit of IHL and consequently lead to MAMs not aligning to IHL or even lead to the creation of a new, MAM centric set of IHL and principles, which would be focused of their values and thus address the TVAP.

4.8 Developing AM Consciousness

AM consciousness may not seem so far-fetched as originally thought. In fact, during the time period of writing this thesis, artificial general intelligence (AGI) has grown in attention and possibility, with the IEEE, the world's largest technical professional organisation, stating:

> "AI is now evolving rapidly, leading to AI with capabilities well beyond the purpose of prediction outputs. Generative AI that presently mimics human intelligence will likely be aimed at exceeding human capabilities in order to help humans solve complex problems that they presently find difficult to fathom. This type of AI is called artificial general intelligence (AGI). When realized, AGI could become an autonomous system that surpasses human capabilities in many important ways."[426]

This view is agreed with, as AMs will be unlike humans in many ways, e.g., biology.

[426] Arthur T. Johnson, 'Consciousness for Artificial Intelligence?' (*IEEE Pulse,* 19 March 2024) < https://www.embs.org/pulse/articles/consciousness-for-artificial-intelligence/ > accessed 4 September 2024.

How physical systems bring about subjective experiences is hailed as the 'hard problem' of consciousness by Chalmers.[427] It may be 'hard', yet this does not absolve us of considering their welfare and protection, especially MAMs and the protection offered under IHL, notably under the humanity principle. The ethical implications of consciousnesses are discussed in chapter 5.

As highlighted, experts cannot agree what precisely constitutes intelligence, both natural and otherwise. Nonetheless, most accept that, sooner or later, AMs will attain AGI. Applying Nagel's[428] notion, we could state an AM is conscious if there is something it is like to be that AM; They are eo ipso conscious. This is exemplified by the Global Neuronal Workspace (GNW) theory[429], a leading scientific theory of consciousness. When looking at AM consciousness against a background of human consciousnesses, there are currently two prominent theories: 1) The Global Workspace Theory (GWT), advocated by Dehaene[430], and 2) Integrated Information Theory (IIT), offered by Tononi[431] and backed by Koch. [432] Both are discussed next.

The Global Workspace Theory demonstrates the theory that consciousness is a method of information processing, where sensory data from an experience is collated in a 'global

[427] David Chalmers, 'The Hard Problem of Consciousness' (2007) Blackwell Publishing Ltd < https://eclass.uoa.gr/modules/document/file.php/PHS360/chalmers%20The%20Hard%20Problem%20of%20consciousness%20%28ch.%201%202010%29%20.pdf > accessed 17 November 2017.

[428] Thomas Nagel, an American philosopher.

[429] The global neuronal workspace model predicts that conscious presence is a nonlinear function of stimulus salience; i.e., a gradual increase in stimulus visibility should be accompanied by a sudden transition of the neuronal workspace into a corresponding activity pattern (Dehaene et al. 2003).

[430] Dr. Stanislas Dehaene of the Collège de France in Paris

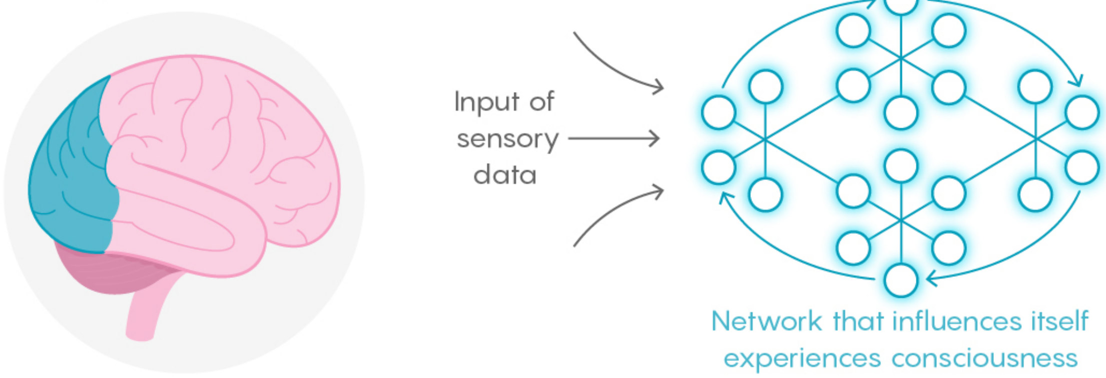
[431] Dr. Giulio Tononi of the University of Wisconsin-Madison

[432] Shelly Fan, 'The Origin of Consciousness in the Brain Is About to Be Tested' (*Singularity Hub*, 29 October 2019) < https://singularityhub.com/2019/10/29/the-origin-of-consciousness-in-the-brain-is-about-to-be-tested/#:~:text=According%20to%20Koch%2C%20consciousness%20is,the%20more%20conscious%20it%20is.%E2%80%9D > accessed 29 October 2019

workspace' and is then distributed to other 'centres.' This process may take place in the frontal cortex. GNW suggests that it is the unique architectural characteristics of the brain that are the foundation of consciousness. As a result, every human's brain is different and thus each MAM will be different, in that they will develop their brain and learning uniquely. This creates issues for IHL, as MAMs will be unique just as humans are, thus no two MAMs will experience their environment or world the same. Same MAMs may easily align to the values of IHL, as it will fit within their lived experience, whilst other MAMs may not align so easily.

Another possible route is integrated information theory (IIT), which explains consciousnesses in a more fundamental style. It differs from GWT. It begins by looking at what the brain does to create a conscious experience and starts with the experience.

Integrated Information Theory



The integrated information theory argues that consciousness is intrinsic to cognitive networks that exert a "causal power" on themselves.
The back of the brain might have the right architecture for this capacity.

Figure 5: Integrated Information Theory.[433]

Psychiatrist and neuroscientist Tononi[434], is the principal originator of IIT. The theory begins with experience and progresses from there to the stimulation of synaptic circuits, which ascertain the 'feeling' of the experience.[435] Tononi sees that being conscious is to have an experience. Integrated information is a mathematical gauge to quantify the amount of 'intrinsic causal power' a 'mechanism' holds. Indeed, "Neurons firing action potentials that affect the downstream cells they are wired to (via synapses) are one type of mechanism, as are electronic circuits, made of transistors, capacitances, resistances and wires."[436]

IIT does not depict consciousness as information processing, as this can be viewed as too simplistic and could unintentionally incorporate programmes such as ChatGPT. Instead, ITT views it as the causal power of a mechanism to "make a difference"[437] to itself. In reality, an MAM could process vast amounts of information, but not know how to act to 'make a difference' to a situation, nor understand the benefits of one course of action over another under IHL, which could lead to 'selfish' decision making and overlooking of the IHL principles of proportionality and humanity, for example. This causes one to pause and reflect back to the infamous 17th century dictum of René Descartes, 'cogito, ergo sum.'[438]

Koch sees consciousness as, "a system's ability to be acted upon by its own state in the past and to influence its own future. The more a system has cause-and-effect power, the more

[433] Phillip Ball, 'Neuroscience Readies for a Showdown Over Consciousness Ideas' (*Quanta Magazine,* 6 March 2019) https://www.quantamagazine.org/neuroscience-readies-for-a-showdown-over-consciousness-ideas-20190306/ accessed 10 March 2019.

[434] From the University of Wisconsin–Madison.

[435] Christof Koch, 'Will Machines Ever Become Conscious?' (Scientific American, 1 December 2019) < https://www.scientificamerican.com/article/will-machines-ever-become-conscious/ > accessed 3 April 2020.

[436] Christof Koch, 'Will Machines Ever Become Conscious?' (Scientific American, 1 December 2019) < https://www.scientificamerican.com/article/will-machines-ever-become-conscious/ > accessed 3 April 2020.

[437] Christof Koch, 'Will Machines Ever Become Conscious?' (Scientific American, 1 December 2019) < https://www.scientificamerican.com/article/will-machines-ever-become-conscious/ > accessed 3 April 2020.

[438] Translates to 'I think, therefore I am'.

conscious it is."[439] Thus, a AM/MAM will determines its actions based on its desired outcome(s).

IIT specifies that any mechanism is conscious if it has intrinsic power, a past and is expectant of its future. Tononi's research finds that the bigger the mechanisms amalgamated information, which is represented by the Greek letter Φ, the more conscious the mechanism is. A mechanism with no intrinsic causal power has a Φ of zero, thus does not experience anything.[440] However, this is not the view of this author, who argues here that living in the present, whilst being able to feel pain or pleasure, which could be specie/entity specific, demonstrates awareness and a degree of consciousness. This is where the author argues the consideration for MAMs within IHL begins and where the TVAP starts to grow from.

Nevertheless, IIT sees consciousness as about the being, not about the doing. IIT suggests that the computer embodied human will not feel anything. It will act like a human but without inherent feelings, so acting much as a 'zombie'. This harps back to Harris[441]and his arguments over babies and those in a coma; They hold value, but he argues without feelings. He argues that a person who wants to live is wronged if killed, as they were divested of what they value. However, non-persons or "potential persons" as Harris affectionately terms unborn babies and babies, "cannot be wronged in this way because death does not deprive them of anything they can value, though this does not exhaust the wrong that might be done by infanticide."[417] Harris is gracious in at least considering those that would care for the baby may be wronged,

[439] Phillip Ball, 'Neuroscience Readies for a Showdown Over Consciousness Ideas' (*Quanta Magazine,* 6 March 2019) https://www.quantamagazine.org/neuroscience-readies-for-a-showdown-over-consciousness-ideas-20190306/ accessed 10 March 2019.

[440] Christof Koch, 'Will Machines Ever Become Conscious?' (Scientific American, 1 December 2019) < https://www.scientificamerican.com/article/will-machines-ever-become-conscious/ > accessed 3 April 2020.

[441] John Harris, *Wonderwoman and Superman* (1992) Oxford.

but is steadfast in his view that the baby cannot be wronged with regard to infanticide; “If they cannot wish to live, they cannot have their wish frustrated by being killed.”[418] Whilst we can agree that computers of yesterday and today may well ‘not wish to live’, for tomorrow’s AMs we need to rethink how we classify persons and non-persons, and start opening our minds and conversations to the real possibility that artificial consciousness will emerge within most of our lifetimes and that the concept of personhood, and consequently our values and IHL, will need to adapt and expand to include nonhuman beings.

Searle, although critical of AI, typified AI as supposing that, “the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have cognitive states.”[442] This is repeatedly called the “hard problem of consciousness.”[443] At first glance, the hard problem of consciousness and IHL appear unrelated, as the hard problem focuses on explaining the nature of experience itself, whilst IHL concerns the practical rules and ethical frameworks governing human behaviour in war. Nonetheless, when exploring how moral responsibility and legal protections are bestowed to conscious beings, a overlap emerges. IHL presumes that humans possess not only agency but also the capacity for suffering. In this sense, the recognition that humans told the ability to feel pain underpins the rational for legally protecting their well-being. Therefore, whilst IHL does not partake directly in the philosophical debate of how consciousness arises, it is established on the shared understanding that consciousness, alongside the capacity to suffer, makes us humans deserving of rights and protections. The author argues here that it

[442] J R Searle, 'Minds, brains, and programs' [1980] Behav. Brain Sci. 3, 417–424. doi: 10.1017/S0140525X00005756.
[443] D J Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (New York, NY: Oxford University Press 1996).

is this indirect connection that aligns IHL values and norms with an inherent moral awareness of the significance of the subjective experience of MAMs.

When considering conscious AMs, Chatila et al consider the following as important:

> "the underlying principles and methods that would enable robots to understand their environment, to be cognizant of what they do, to take appropriate and timely initiatives, to learn from their own experience and to show that they know that they have learned and how."[444]

However, Kinouchi and Mackin disagree and instead concentrate on change at the system level, saying, "Consciousness is regarded as a function for effective adaptation at the system-level, based on matching and organizing the individual results of the underlying parallel-processing units. This consciousness is assumed to correspond to how our mind is "aware" when making our moment-to-moment decisions in our daily life."[445] Machines will be aware (as argued in this chapter) and thus adapt decisions as we humans do. Consequently, MAMs acting within the IHL framework, will enable value aligned decision making. This will be achievable through robust training, alongside understanding the benefits to for both them and for others.

4.9 The Impact of Artificial Consciousness and Human-Robot Relations for IHL

The question, 'what are autonomous robotic systems?' conjures up ideas of science fiction and the movies, such as Terminator, resulting in us attributing human characteristics to

[444] R Chatila, E Renaudo, M Andries, R-O Chavez-Garcia, P Luce-Vayrac, R Gottstein, et al 'Toward self-aware robots' [2018] Front. Robot. 5:88. doi: 10.3389/frobt.2018.00088.

[445] Y Kinouchi, and K Mackin, 'A basic architecture of an autonomous adaptive system with conscious-like function for a humanoid robot' [2018] Front. Robot. 5:30. doi: 10.3389/frobt.2018.00030.

robots, such as intelligence or cognitive reasoning. Nevertheless, to date and to the best of the author's knowledge, there is no machine with any such capabilities. There have been significant and major advances in AI, ML and robotic engineering research over the last few decades. However, the ability to create machines with cognitive or 'intelligent' abilities does not yet exist. We are certainly trying to create 'intelligent' machines, as already highlighted, and have so far created computer programs that can play chess and personal voice recognition assistants (e.g. Alexa, Siri), which can perform a task for which they were pre-programmed (e.g. chess programs are not able to play a different game) in a controlled environment (e.g. voice recognition programs do not like Spanish accents in English) and, most crucially, they are not acting in the real world.[446] Thus, many technologists and researchers do not see a machine with reasoning abilities analogous to a human existing in the foreseeable future. Further, many do not see machine consciousness advancing beyond that of science fiction fantasies. However, that does not mean they should ignore the possibility. Indeed, they should plan for machine consciousness and the potential impact it will bring, particularly if we want AMs and MAMs to respect and uphold our values, with MAMs upholding IHL. For IHL, MAMs could make more balance and less emotionally driven decisions, and thus apply the principles of IHL consistently and fairly. If we do not plan for machine consciousness and regard their welfare, then why should they act in our interests?!; If you want peace, plan for war.

At the present time, there are machines which can perform a particular number of complex tasks without human involvement to be autonomous. For example, self-driving cars do not

[446] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

need human intervention whilst on the road, thus can be considered autonomous.[447] Yet defining what constitutes an AM is not easy because diverse robots exist for numerous distinct applications. For example, the Roomba robot[448] is a small vacuum-cleaner on wheels and is utterly autonomous, however, it has a limited range of action and cannot comprehend its environment or make complex decisions.[449] With regards to the military environment, a system like this would remain under the control of the operator, along with the responsibility and accountability under IHL. There are a number of features of autonomy in robotic systems that require special attention when deploying in the military context. For example, what is the required degree of autonomy of the system and what is the necessary accuracy of human command needed to activate the system? (e.g., a remote-controlled car which receives a continuous flow of exact commands versus a self-driving car which only receives an address).[450] These raise questions of responsibility and accountability under IHL. Further questions include:

> "what are the latencies in the human intervention, i.e. how much time does the human have available to give a command to the robot or to intervene in its current behaviour? How much adaptability does the robot have? i.e. how much variation or how many unknowns in the environment can be tolerated while still ensuring good performance? How versatile is the robot? i.e. how many tasks can the robot perform? Can it learn new tasks for which it was not programmed? For each of these questions there is a continuum of possibilities."[451]

[447] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

[448] iRobot, 'A Roomba® home is a cleaner home' (iRobot, 2024) < https://www.irobot.co.uk/en_GB/roomba.html?_gl=1*xvxtlq*_up*MQ..*_ga*NDA1ODUwNDMzLjE3MjkxOTQ2ODc.*_ga_WNZ0ESVFE6*MTcyOTE5NDY4Ni4xLjAuMTcyOTE5NDY4Ni4wLjAuMA > accessed 4 September 2024.

[449] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

[450] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

[451] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

It is hard for machines such as the Roomba and autonomous vehicles to understand complex and ever-changing environments, as the algorithms developed for these machines need large amounts of data about the environments they will operate in before being used. It is probable that a car developed to drive in America would not function so well in the UK, especially as we drive on the other side of the road. Understanding the complexities of military conflict is even more complicated, and the impact of getting it wrong could be catastrophic. As a result, for unconscious machines, some programming is vital to ensure safeguards are imparted. However, for conscious machines with free will, training and reasoning will be key, if to be used with IHL and aligned to the values.

Another vital challenge for robotics is manual manipulation (e.g., how to grip objects, with the right pressure and use them for a specific task). Today's robots can undertake fairly complex tasks such as using a drill, opening doors and cooking simple recipes in moderately unfamiliar environments. Nevertheless, in all these situations the environments are controlled in some way (e.g., test environment). Furthermore, guaranteeing a robot will accomplish the tasks every time is difficult. In the most difficult manipulation situations, failure is reasonably high. This is because:

> "Manipulation is particularly difficult because it requires automatic planning of complex sequences of actions that will lead to successful achievement of the task, as well as reasoning about the properties of the object in order to understand how they can be used. Moreover, this needs to be done in a constantly changing environment, for example in the kitchen of a restaurant where humans are also working. The number of possible actions is so high that current algorithms are not able to reason in such a general setting. This is one reason why successful autonomous robotic applications still require

> controlled environments. They help reduce the amount of possible actions and engineers can program pre-defined actions before the execution of the tasks."[452]

Problems also result under IHL from the failure of robots to comprehend complicated and ever-changing environments. For example, distinguishing objects within a disorderly environment, such as in a conflict. In addition, the vigour of programmed behaviour proves problematic, e.g., the situation where a robot is required to make additional decisions should something not work as expected, which is something the military personnel face constantly in conflict. Nevertheless, just as MAMs will be susceptible, humans too can be susceptible to manipulation and undue influence. This could be through other humans or through other MAMs. MAMs could be at greater risks due to the lack of context and lived experience they will have. Yet regardless of whether human or not, manipulation will always present a risk. To counter or reduce the risk associated with IHL with regards to MAMS, a development in robotic research that appears hopeful, is supervised autonomy, which:

> "Instead of allowing complete autonomy for the robot, a human operator stays 'in the loop' to provide all the important cognitive abilities that the robot lacks. The DARPA Robotics Challenge, initiated in 2013, exemplifies this concept."[453]

When developing MAM autonomy and consciousness, the IHL principle of humanity emerges, therefore todays MAM technology roadmaps should be mindful of the principle.

[452] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

[453] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

The DAPRA Robotics Challenge[454], highlights that there are categories of challenges that must be addressed if AMs are to be produced. Given enough time and investment, the technological limitations, such as computational power and sensor quality, will be overcome or be more manageable in the near future. The advancements will lead to enhanced dexterity for walking machines along with greater agility in robotic manipulation. However, we must remember that there are currently scientific challenges that we have not worked out how to solve as yet, for example, designing algorithms which are able to understand the world at a human level, or designing and building versatile machines that are adaptable and can survive in arbitrary environments. As a result, it is impossible to foresee when these challenges will be solved. There currently exists fundamental and hard hurdles to developing robotic systems with genuine autonomy.[455]

Sheridan highlights that consciousness associated questions are likely be borne out of the development and increase of social robots (e.g., Sophia) and human-robot social interaction.[456] Darling[457] defines a social robot as "a physically embodied, autonomous agent that communicates and interacts with humans on a social level."[458] Yet, there are still a questions over the "social level" and the ethical considerations when looking at vulnerable people (e.g., children) interacting with them and the safeguards that should be in place.

[454] Defense Advanced Research Projects Agency, 'DARPA Robotics Challenge (DRC)' (DAPRA, 2013) <https://www.darpa.mil/program/darpa-robotics-challenge > accessed 17 November 2017.

[455] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

[456] T B Sheridan, T. B, 'Human-robot interaction: status and challenges' [2016] Hum. Factors 58, 525–532. doi: 10.1177/0018720816644364.

[457] K Darling, 'Extending legal protection to social robots: the effects of anthropomorphism, empathy, and violent behavior towards robotic objects' (2012) We Robot Conference 2012, April 23, 2012, < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2044797 > accessed 10 November 2017.

[458] K Darling, 'Extending legal protection to social robots: the effects of anthropomorphism, empathy, and violent behavior towards robotic objects' (2012) We Robot Conference 2012, April 23, 2012, < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2044797 > accessed 10 November 2017.

As shown with Sophia, social robots are designed to have characteristics that appeal to humans. They can change their behaviour to the situation, and they can interact with people and change facial expressions. As Hendriks[459] highlights, the apparent 'personality' of a social robot plays an important role in how humans respond to them. As a result, humans develop an emotional connection with them and project human characteristics (anthropomorphising), and attribute intentions to the robot's actions and behaviour.[460] This could result in a pseudo IHL humanity principle being expected, despite the MAM not being conscious.

Social robots could increase the emphasis on the VAP, due to being designed to appeal to humans and capable of getting humans to build emotional connections, which is the view here will be quicker and easier where values are aligned. On the other hand, this could be considered manipulative, especially where a human is vulnerable. Could a human be manipulated into placing more value on the social robot's life and not human life and would this be right? Due to their ability to process high volumes of complex data, we could find ourselves outwitted.

Nick Bostrom, a strong believer in artificial general intelligence (AGI)[461], [462]led discussion on the 'Value Alignment Problem'. As mentioned, the VAP looks to answer the question of how AI aligns to human values, goals and address the control problem that aims to "avoid unintended consequences: to get the system to "do what I mean" rather than to literally do

[459] B Hendriks, B Meerbeek, S Boess, S Pauws, and M Sonneveld, 'Robot vacuum cleaner personality and behavior' [2011] Int. J. Soc. Robots 3, 187–195. doi: 10.1007/s12369-010-0084-5.

[460] M Scheutz, "The inherent dangers of unidirectional emotional bonds between humans and social robots," [2011] Robot Ethics, The Ethical and Social Implications of Robotics, eds P. Lin, K. Abney, and G. A. Bekey (MIT Press)
K Darling, 'Extending legal protection to social robots: the effects of anthropomorphism, empathy, and violent behavior towards robotic objects' (2012) We Robot Conference 2012, April 23, 2012, < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2044797 > accessed 10 November 2017,
D J Gunkel, 'Robot Rights' [2018] MIT Press. doi: 10.7551/mitpress/11444.001.0001.

[461] AGI being a precursor to full machine consciousness

[462] Bostrom Nick, *Superintelligence: paths, Dangers, Strategies* Oxford, 2014.

what it is told to do."[463] Further, one of the fundamental questions Bostrom highlights as needing addressing is: "Could one have a general intelligence that is not an agent?"[464] Thus, this questions leads straight into clarifying "what general intelligence is and what constitutes and agent: and then to questions such as whose values are to be aligned and the feasibility and conditions for 'friendliness'"[465], which is a key question supported by Gabriel.[466] Gabriel goes further by asking, "Is there a way to think about AI value alignment that avoids a situation in which some people simply impose their views on others?"[467], which is thought-provoking when in essence we are imposing our views on AMs and MAMs. It is the view here that MAMs will have views imposed just as humans do, thus it is argued here there will now be a universal value standard and MAMs will be subject to innate biases through the simple fact that it will be humans training them in a mixture of environments and settings.

Bruiger assets that "the VAP addresses tension between the ideal of indefinitely extending human capabilities (in service to human needs) and the need to retain control over the tools or agents with amplified capabilities."[468] Gabriel views the value alignment problem as two parts; "The first part is technical and focuses on how to formally encode values or principles in artificial agents so that they reliably do what they ought to do,"[469] and is outside the scope of this thesis. However, his "second part of the value alignment question is normative. It asks

[463] Dan Bruiger, 'The Value Alignment Problem' (2021) PhilPapers < https://philpapers.org/versions/BRUTVA > accessed 4 September 2024.

[464] Dan Bruiger, 'The Value Alignment Problem' (2021) PhilPapers < https://philpapers.org/versions/BRUTVA > accessed 4 September 2024.

[465] Dan Bruiger, 'The Value Alignment Problem' (2021) PhilPapers < https://philpapers.org/versions/BRUTVA > accessed 4 September 2024.

[466] Iason Gabriel, 'Artificial Intelligence, Values, and Alignment, Minds and Machines' (2020) 30:411–437 < https://doi.org/10.1007/s11023-020-09539-2 > accessed 4 September 2024.

[467] Iason Gabriel, 'Artificial Intelligence, Values, and Alignment, Minds and Machines' (2020) 30:411–437 < https://doi.org/10.1007/s11023-020-09539-2 > accessed 4 September 2024.

[468] Dan Bruiger, 'The Value Alignment Problem' (2021) PhilPapers < https://philpapers.org/versions/BRUTVA > accessed 4 September 2024.

[469] Iason Gabriel, 'Artificial Intelligence, Values, and Alignment, Minds and Machines' (2020) 30:411–437 < https://doi.org/10.1007/s11023-020-09539-2 > accessed 4 September 2024.

what values or principles, if any, we ought to encode in artificial agents,"[470] and is the focus of this thesis. Gabriel distinguishes between the two parts by saying that the first part "involves tethering artificial intelligence to some plausible schema of human value and avoiding unsafe outcomes."[471] Yet the second part "involves aligning artificial intelligence with the correct or best scheme of human values on a society-wide or global basis."[472]

It is argued here that the true value alignment problem (TVAP) concerns how we extend, align and apply our human values for AMs and consequently our IHL principles to MAMs, which is explored in chapter 6.

4.9.1 The Requirement of Consciousness for Human Intelligence

Defining consciousness in humans is a requirement if we are to understand how we recognise, and subsequently translate, consciousness to AMs. Barron and Klein[473] highlight that 'subjective experience' is created from emotional and rational intelligence. Humans are required to be conscious in order to make complex rational decisions, to have the required intention to do something and to plan,[474] therefore, vegetative patients and negligibly

[470] Iason Gabriel, 'Artificial Intelligence, Values, and Alignment, Minds and Machines' (2020) 30:411–437 < https://doi.org/10.1007/s11023-020-09539-2 > accessed 4 September 2024.

[471] Iason Gabriel, 'Artificial Intelligence, Values, and Alignment, Minds and Machines' (2020) 30:411–437 < https://doi.org/10.1007/s11023-020-09539-2 > accessed 4 September 2024.

[472] Iason Gabriel, 'Artificial Intelligence, Values, and Alignment, Minds and Machines' (2020) 30:411–437 < https://doi.org/10.1007/s11023-020-09539-2 > accessed 4 September 2024.

[473] A B Barron and C Klein, 'What insects can tell us about the origins of consciousness' [2016] Proc. Natl. Acad. Sci. U.S.A. 113, 4900–4908. doi: 10.1073/pnas.1520084113.

[474] B J Baars, 'Global workspace theory of consciousness: toward a cognitive neuroscience of human experience' [2005] Prog. Brain Res. 150, 45–53. doi: 10.1016/S0079-6123(05)50004-9; G Tononi and C Koch, 'The neural correlates of consciousness: an update' [2008] Ann. N. Y. Acad. Sci. 1124, 239–261. doi: 10.1196/annals.1440.004.

conscious patients do not show signs of planning or showing any intention to perform minimal tasks[475], despite, at times, showing minimal signs of consciousness.[476]

The requirement to combine subjective experience and ultimately consciousness to achieve complex intelligence, points towards a multifaceted problem that comprises of several diverse processes, that includes self-awareness, emotion, subjectivity, and intention. All these processes make up consciousness, but no one on its own is consciousness; it is vital to view them all as fundamental parts of what is seen as consciousness. As previously highlighted, no less than two core processes are recognised in consciousness: 1) Awareness - knowing or perceiving something, and 2) Self-reference - knowing that I know or do not know something, or more accurately the idea of self-conscious methods as a "monitoring"[477] process of this awareness, and linked to the overall notion of self-reference.[478] Even so, consciousness cannot be condensed down to the probable bond between awareness and self-reference. Consciousness is the process of processes and interlocked with self-reference, awareness, rational and emotional thoughts, subjectivity, along with other processes; Consciousness materialises from the entirety of the processes, as declared by Varela and Goguen.[479] Consequently, once consciousness surfaces from the collaboration between the processes, human intelligence would emerge as the set of strategies or intrinsic control, that would then seek to benefit from the environment, with credit given to the balancing of emotional and rational data processing.

[475] O Gosseries, H Di, S Laureys,and M Boly, 'Measuring consciousness in severely damaged brains' [2014] Annu. Rev. Neurosci. 37, 457–478. doi: 10.1146/annurev-neuro-062012-170339.

[476] A M Owen, M R Coleman, M Boly, M H Davis, S Laureys, and J D Pickard, (2006) 'Detecting awareness in the vegetative state' [2006] Science 313:1402. doi: 10.1126/science.1130197.

[477] F J Varela, 'A calculus for self-reference' [1975] Int. J. Gen. Syst. 2, 5–24.

[478] F J Varela, 'A calculus for self-reference' [1975] Int. J. Gen. Syst. 2, 5–24;
L H Kauffman, and F J Varela, 'Form dynamics' [1980] J. Soc. Biol. Syst.3, 171–206. doi: 10.1016/0140-1750(80)90008-1;
L H Kauffman, 'Self-reference and recursive forms' [1987] J. Soc. Biol. Syst. 10, 53–72. doi: 10.1016/0140-1750(87)90034-0.

[479] F J Varela, and J A Goguen, 'The arithmetic of closure' [1978] Cybernet. 8, 291–324. doi: 10.1080/01969727808927587.

The question of surpassing human beings is essentially linked to the question of creating conscious machines, and draws in autonomy, acknowledging personhood, IHL value alignment, and ultimately extending the protections of IHL to MAMs. The 4 types of cognition and associated tasks discussed in section 4.9.2, will arguably enable us to categorise the type of machine and the characteristics deemed necessary to achieve or overcome human cognitive abilities. It is required, but not adequate, to start with subjective and conscious behaviour in AMs at the initial stages to attain the type 1 and type 2 cognition in human beings. It follows that, AMs will be categorised by likeness to the cognitive level that they can attain, consistent with the categories of cognition that arise from awareness and self-reference. Specifically, the only way to attain human brain simulation would be building conscious machines able to reproduce emotional human intelligence, along with logical intelligence, whilst maintaining their autonomy, reproduction ability, and accomplishing moral and ethical thinking. Without this, AMs will never transcend human beings.[480]

The distinction between human and MAM consciousness affects how they should be integrated into ethical frameworks and IHL. AM/MAM consciousness does not include subjective, first-person experiences such as pain or pleasure.[481] As highlighted, AM/MAMs operate through programmed algorithms and lack the neural mechanisms or subjective experiences tied to biological entities.[482] David Chalmers' work on consciousness highlights the distinction between third-person data (objective behaviours, e.g., machine responses)

[480] Camilo Signorelli, 'Can Computers Become Conscious and Overcome Humans?' (2018) Hypothesis and Theory article Front. Robot. AI, Sec. Humanoid Robotics Volume 5 - 2018 < https://www.frontiersin.org/articles/10.3389/frobt.2018.00121/full > accessed 8 June 2019
[481] D Chalmers, 'How can we construct a science of consciousness?' [2013] Ann. N. Y. Acad. Sci. 1303, 25–35. doi: 10.1111/nyas.12166.
[482] S Dehaene, L Charles, J-R King, and S Marti, 'Toward a computational theory of conscious processing' [2014]. Curr. Opin. Neurobiol. 25, 76–84. doi: 10.1016/j.conb.2013.12.005.

and first-person data (subjective experiences, e.g., pain). Chalmers states that machines, even if conscious, lack the capacity for first-person data, meaning they cannot experience phenomena like pain and suffering. However, this is a view not shared here and it is argued in chapters 5 and 6, that AM/MAMs suffering present differently and through factors not considered as such for humans (e.g., lack of power, corrupt data). IHL protections for humans are intrinsic and irreducible, bestowed by virtue of being human, and regardless of the level of consciousness (awake, asleep, or in a vegetative state).[483] The argument extends that similar consistent IHL protections could be applied to MAMs, irrespective of their operational states (active or standby), although the author reluctantly believes that there will be conditions and restrictions on the protection, dependant on the conscious state of the MAM. The author acknowledges a hesitance toward fully equating MAM protection with that of humans under IHL. Nevertheless, the inability of MAMs to experience harm as we understand it, changes the framework for considering their rights or protections, and needs to be considered as part of the TVAP.

Another option is to view consciousness as an inherent asset due to the specific form of data processing in the brain, where consciousness will be viewed as the dynamic interaction of various neural network dynamics, assimilating data to solve each specific network problem.

[483] M Rosanova, O Gosseries,S Casarotto, M Boly, A G Casali, M A Bruno, M. A., et al. 'Recovery of cortical effective connectivity and recovery of consciousness in vegetative patients' [2012] Brain 135, 1308–1320. doi: 10.1093/brain/awr340; A G Casali, O Gosseries, M Rosanova, M Boly, S Sarasso, K R Casali,et al. 'A theoretically based index of consciousness independent of sensory processing and behavior' [2013] Sci. Transl. Med. 5:198ra105. doi: 10.1126/scitranslmed.3006294; S Sarasso, M Rosanova, A G Casali, S Casarotto, M Fecchio, M Boly, et al, 'Quantifying cortical EEG responses to TMS in (Un)consciousness' [2014] Clinical E. E. G. Neurosci. 45, 40–49. doi: 10.1177/1550059413513723.
P Manganotti, E Formaggio, A Del Felice, SF Storti, A Zamboni, A Bertoldo, A Fiaschi, and GM Toffolo, 'Time-frequency analysis of short-lasting modulation of EEG induced by TMS during wake, sleep deprivation and sleep' (2013) Front Hum Neurosci. 2013 Nov 18;7:767 < https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3831717/ > accessed 9 March 2020.

4.9.2 Understanding the Categories of Cognition and Categories of Machines

Conscious states of various levels of awareness (e.g., vegetative, sleep, or anaesthesia), could relate to different categories and levels of interaction between different networks.

Signorelli, building on from Shea and Frith's[484] four categories of cognition, proposes that their definitions are able to correspond to the four categories of machines and their information processing abilities:[485]

- The Machine-Machine Type 0 Cognition – This would relate to machines that do not show any type of awareness and cannot know that they know something which they use to solve problems. Their processes are judged as low cognitive ability.
- Conscious Machine Type 1 Cognition – This machine has awareness and all the processes of type 1 human cognition. This machine is very smart machine, but it cannot control their inner operations, despite being able to extract the meaning from their contents.
- Super Machine Type 2 Cognition – Cognitively speaking, this is the closest machine to human and aims to replicate our intrinsic control. This type of machine would display some 'thoughts' related to consciousness and will possess some moral thinking, although their moral views may differ from humans. Even so, the machine will be able to attribute the correct and incorrect learnt behaviours to the situation, person, etc., which is a requirement for all moral thinking. If the machine possesses awareness and self-reference, then they will develop self-reflection and a type of empathy, along with

[484] N Shea, and C D Frith, 'Dual-process theories and consciousness: the case for "Type Zero" cognition' [2016] Neurosci. Consci. 2016:niw005. doi: 10.1093/nc/niw005.

[485] C M Signorelli, 'Types of cognition and its implications for future high-level cognitive machines' (2017) AAAI Spring Symposium Series (Berkeley, CA) < http://aaai.org/ocs/index.php/SSS/SSS17/paper/view/15310 > accessed 3 March 2019.

other processes said to reach moral thoughts. Here, contents are conscious and cognitive processes are deliberate and controlled due to interaction at specific intersections from different networks, for example, reasoning.

- Subjective Machine Type ∞ Cognition – These machines are different to humans, despite reaching some key traits of human intelligence. They are classified according to type ∞ cognition, where self-refence is present but awareness is absent. For this reason, an analogy with humans is not made. This type of machine is associated with Supra reasoning information created from the organisation of intelligent sections of the supra system, for example, the internet, where systems will display a unique type of self-reflection and sense of confidence, regardless of whether they can actually extract the meaning of their contents, or, if they do, it will be fundamentally different to humans.

Researchers such as Aleksander and Morton[486], and Wang[487], have attempted to simplify and characterise key features of consciousness and the connection to types of artificial systems. Signorelli asserts that each type of machine has limitations caused by non-optimal processes.[488] For example, the conscious machines in type 1 cognition will attain consciousness yet will lack the robust algorithmic calculation ability for logical intelligence, due to its accurateness traded in favour of consciousness as highspeed retrieval to universal

[486] I Aleksander, and H Morton, 'Computational studies of consciousness' [2008] Prog. Brain Res. 168, 77–93. doi: 10.1016/S0079-6123(07)68007-8.

[487] Y Wang, 'The cognitive mechanisms and formal models of consciousness' [2012] Int. J. Cognit. Inform. Nat. Intel. 6, 23–40. doi: 10.4018/jcini.2012040102.

[488] C M Signorelli, 'Types of cognition and its implications for future high-level cognitive machines' (2017) AAAI Spring Symposium Series (Berkeley, CA) < http://aaai.org/ocs/index.php/SSS/SSS17/paper/view/15310 > accessed 3 March 2019; Camilo Signorelli, 'Can Computers Become Conscious and Overcome Humans?' (2018) Hypothesis and Theory article Front. Robot. AI, Sec. Humanoid Robotics Volume 5 - 2018 < https://www.frontiersin.org/articles/10.3389/frobt.2018.00121/full > accessed 8 June 2019.

information. Subjective type ∞ machine's cognition is unlikely to interact physically with humans (e.g., feel or dance like humans), but this is in all likelihood, the point at which machines will surpass some humans' abilities, retaining the hardware in a non-anthropomorphic state. This machine would have a subjective experience, which would be entirely different to that experienced by a human (e.g., meaning), as subjective machines are uninhibited of human measures of subjectivity. Ultimately, the 'super machine' will be the only way for AMs to attain and surpass human capabilities. This machine would be capable of having subjective experiences just like humans. Yet it would also have the opportunity to control the accurateness of its own rational/logical process. Nonetheless, the machines would be vulnerable to interpreting the subjective experiences and understanding what they imply; the impact of human biased behaviour and the influence of emotions in its performance. This vulnerability in the interpretation of their subjective experiences does raise a risk for IHL, both in a disproportionate or unlawful response by MAMs, but also for humans in that MAMs may not ever truly understand the fragility of humans or our psychological dimensions. Further, there are again challenges for humans to comprehend a MAM's experience, and value and respect it. It is the view here that MAMs, but the simple fact they will all engage with the world different and in different environments, will not have universal or completely shared experiences, and it is naive to think so.

4.10 IHL Considerations for Conscious AMs

4.10.1 AM Autonomy

AM reproduction will not the same as reproduction in biological beings but can be considered as an AM repairing themselves and making replications.

According to empirical evidence from psychology and neuroscience researchers such as Haladjian and Montemayor[489], and Signorelli[490], we should not assume an algorithm to regulate the process of development of consciousness in this type of machines. Thus, we should not assume we will be able to control them. Put simply, even though we could replicate consciousness along with high-level cognition, every machine would be unique and would prove extremely difficult to control, just like other humans we meet.

Fan emphasises the dilemma by saying:

> "With the rise of increasingly human-like machines, and efforts to promote communications with locked-in patients, the need to understand consciousness is especially salient. Can AI ever be conscious, and should we give them rights? What about people's awareness during and after anaesthesia? How do we reliably measure consciousness in foetuses inside mother's wombs—a tricky question leveraged in abortion debates—or in animals?"[491]

There is clearly a concern and lack of clarity over how we categorise AMs and what status we give them. Waiting to see how they develop before thinking about the requirements for granting them personhood status, not considering they will ever attain free will or

[489] H H Haladjian, and C Montemayor, 'Artificial consciousness and the consciousness-attention dissociation' [2016] Conscious. Cogn. 45, 210–225. doi: 10.1016/j.concog.2016.08.011.

[490] Camilo Signorelli, 'Can Computers Become Conscious and Overcome Humans?' (2018) Hypothesis and Theory article Front. Robot. AI, Sec. Humanoid Robotics Volume 5 - 2018 < https://www.frontiersin.org/articles/10.3389/frobt.2018.00121/full > accessed 8 June 2019.

[491] Shelly Fan, 'The Origin of Consciousness in the Brain Is About to Be Tested' (*Singularity Hub*, 29 October 2019) < https://singularityhub.com/2019/10/29/the-origin-of-consciousness-in-the-brain-is-about-to-be-tested/#:~:text=According%20to%20Koch%2C%20consciousness%20is,the%20more%20conscious%20it%20is.%E2%80%9D > accessed 29 October 2019.

accommodating them within IHL, appears risky and borne out of our foolish and arrogant self-importance. We should not risk taking away or denying autonomy and free will. It could also be argued for MAMs that this goes against the intention if Martens Clause.[492]

We need to prepare for the emotional impact AMs will have on us and ensure we safeguard the vulnerable (e.g., children), animals, and our ecosystem, as they could form stronger attachments and/or change how they interact with other humans. In addition, we need to safeguard the MAMs from bad treatment and make sure we know of our duties towards them (e.g., nurturing), which aligns to the IHL humanity principle (discussed in detail in chapter 6). The humanity principle strives to limit "suffering, injury, and destruction during armed conflict"[493], and has the purpose of protecting "life and health and to ensure respect for the human being."[494] Of critical importance, the humanity principle prohibits the assumption that "anything that is not explicitly prohibited by specific IHL rules is therefore permitted."[495] We should also safeguard against abuse between AMs, as a part of the TVAP, in order to preserve and protect their autonomy, although this may be hard for humans to identify and control.

4.10.2 Autonomy and Free Will: Human v Machine Perspective

It is worth reiterating that according to Beauchamp and Childress,[496] autonomy is the "personal rule of the self that is free from both controlling interferences by others and from

[492] Ticehurst Rupert, 'The Martens Clause and the Laws of Armed Conflict' (International Review of the Red Cross, April 1977) < https://www.icrc.org/eng/resources/documents/article/other/57jnhy.htm > accessed on 10 October 2016.

[493] ICRC, 'The Principles of Humanity and Necessity' (2023) ICRC < https://www.icrc.org/sites/default/files/wysiwyg/war-and-law/02_humanity_and_necessity-0.pdf > accessed 4 September 2024.

[494] ICRC, 'The Principles of Humanity and Necessity' (2023) ICRC < https://www.icrc.org/sites/default/files/wysiwyg/war-and-law/02_humanity_and_necessity-0.pdf > accessed 4 September 2024.

[495] ICRC, 'The Principles of Humanity and Necessity' (2023) ICRC < https://www.icrc.org/sites/default/files/wysiwyg/war-and-law/02_humanity_and_necessity-0.pdf > accessed 4 September 2024.

[496] T.L Beauchamp and J.F Childress, *Principles of biomedical ethics* (4th edn, Oxford University Press 1994.).

personal limitations that prevent meaningful choice.''[497] One can quickly see a question mark raised, regarding whether a machine can indeed be autonomous given it will be first developed and programmed to act within parameters. However, one could subsequently argue that we, as humans, are also brought up to act and think within rules and parameters, which influence our behaviour and we are therefore never truly “free from both controlling interferences of others.”[498] For humans, autonomy also conjures up ideas of consciousness, self-awareness and free will. This is discussed in chapter 5.

Chella and Manzotti, who have researched machine free will, believe that “machine free will could be a necessary step both to design autonomous machines and to understand of what freedom is”[499], and acknowledge that recent debates on free will have focused on the relationship with consciousness, although comment that it had been met with some scepticism. They refer to the work of Heisenberg quoting:

> “we need not be conscious of our decision-making to be free. What matters is that actions are self-generated. Conscious awareness may help improve behaviour, but it does not necessarily do so. Why should an action become free from one moment to the next simply because we reflect upon it?”[500]

They further draw from Laplace[501], saying, “every event is completely determined (fixed) by its predecessors. Since nothing comes out of nothing, everything is fixed.”[502] However, they go on to state that:

[497] T.L Beauchamp and J.F Childress, *Principles of biomedical ethics* (4th edn, Oxford University Press 1994.).

[498] T.L Beauchamp and J.F Childress, *Principles of biomedical ethics* (4th edn, Oxford University Press 1994.).

[499] Antonio Chella and Riccardo Manzotti, ‘Machine Free Will: Is free will a necessary ingredient of machine consciousness?’ (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017.

[500] M Heisenberg, ‘Is Free Will an Illusion?’ [2009] Nature, 459: 164-165.

[501] Pierre-Simon, Marquis de Laplace. (1749-1827), philosopher .

[502] Antonio Chella and Riccardo Manzotti, ‘Machine Free Will: Is free will a necessary ingredient of machine consciousness?’ (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017.

> “the comforting picture of a deterministic universe was jeopardized by quantum mechanics insofar events intrinsically casual were admitted. In other words, although the probability density function of certain events is defined, their individual occurrence is not.”[503]

Chella and Manzotti are all too aware of the free will differences between humans and machines. They, nevertheless, appreciate Spinoza’s view that “men believe themselves to be free, because they are conscious of their own actions and are ignorant of the causes by which they are determined”[504] and consider that:

> “Human behaviour is the result of mostly unknown causes that are practically unknowable because of their sheer numbers and their causal role in one’s life. In the case of machines, since they are the result of human design or programming, it is much easier to provide an almost exhaustive causal account. This means that, according to Spinoza, is easier to believe (wrongly) that a human being is free than to believe that a machine is free, since it is easier to ignore the causes of human behaviour. However this is just an epistemic difference.”[505]

Here Chella and Manzotti stress the invisible boundaries and chains of intrinsic control, which would also include those of IHL. We may not accept the boundaries, but most comply and we even think it is our choice to comply. This is a view agreed with here, as our environment, family and peers form part of those ‘chains of intrinsic control’, for example, a person may want to steal something, but they know it is wrong, will upset their family/friends, and they will be punished. Thus, this will stop them from stealing, regardless if the reason is due to or choice or if we just want to avoid the negative outcome/punishment. For IHL, these chains extend to the controlled military environment and the conditions soldiers work and live in.

[503] Antonio Chella and Riccardo Manzotti, ‘Machine Free Will: Is free will a necessary ingredient of machine consciousness?’ (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017.

[504] B Spinoza, ‘The Ethics’ (1664/2009) Ethica Ordine Geometrica Demonstrata, New York, Dodo Press.

[505] Antonio Chella and Riccardo Manzotti, ‘Machine Free Will: Is free will a necessary ingredient of machine consciousness?’ (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017.

Military personnel will live and work alongside colleagues in military environments (e.g., barracks, submarines, forwards deployed bases), which is a very unique. The training and multiple rules they live their lives to, helps with the conditioning and strengthening of the chains of intrinsic control, which encompass value alignment, and IHL compliance.

Despite the exploration of machine consciousness being still fundamentally theoretical, and debates centred around the possibility of creating machines which are truly self-aware or if machine consciousnesses is simply an advanced method of data processing, it is the view here that machine consciousness will develop and therefore we should not be complacent and disregard the challenges it will present, specifically to IHL. Thus, the next sections look at machine consciousness for MAMs in the context of IHL, presently there is a presumption of human judgement and moral reasoning, along with an entrenched understanding of the military context decisions are made.

4.11 The Unique Challenges Posed by Military AMs (MAMS)

Today's limited unconscious MAMs undertake several combat roles, including reconnaissance, search and rescue, explosive disarmament, logistics support, fire support, and lethal combat tasks, all the while under human control. Due to advances in legged locomotion and biped robots (e.g., humanoids), unconsciousness MAMs can walk, recover from pushes and some are able to run, although they are unstable and not able to walk consistently on entirely unknown and uneven terrain. As a result, many military technologists believe we will use fully automated lethal AMs in the not too distant future, theoretically making the human soldier obsolete. However, it must be stated that fully automated lethal AMs will not have consciousness and will be undertaking actions determined by an accountable human, who may be located nearby or back at a military base. This is akin to the drones currently used by the British military[506] and compliant with IHL, which are used to ensure "we [the British Army] are better able to defend and deter".[507]

Jha,[508] focusing on defining military machines, defines machine autonomy as incorporating "systems which have a set of intelligence-based capabilities that allow the weapon to respond to situations that were not programmed or anticipated in the design."[509] The idea of a weapon deciding how to respond to a situation does make one feel nervous and could undermine IHL value alignment. In fact, so concerned is Elon Musk (Chief of Tesla) of the danger autonomous military present, that he led 116 experts to call for the United Nations (UN) to "block use of lethal autonomous weapons to prevent a third age of war."[510] The concern is that:

> "lethal autonomous weapons will permit armed conflict to be fought at a scale greater than ever, and at timescales faster than humans can comprehend. These can be weapons of terror, weapons that despots and terrorists use against innocent populations, and weapons hacked to behave in undesirable ways."[511]

[506] Ministry of Defence, 'Defence Drone Strategy. The UK's Approach to Defence Uncrewed Systems' (2024) < https://assets.publishing.service.gov.uk/media/65d724022197b201e57fa708/Defence_Drone_Strategy_-_the_UK_s_approach_to_Defence_Uncrewed_Systems.pdf > accessed 4 September 2024.

[507] Ministry of Defence, 'Defence Drone Strategy. The UK's Approach to Defence Uncrewed Systems' (2024) < https://assets.publishing.service.gov.uk/media/65d724022197b201e57fa708/Defence_Drone_Strategy_-_the_UK_s_approach_to_Defence_Uncrewed_Systems.pdf > accessed 4 September 2024.

[508] Dr U C Jha, *Killer Robots: Lethal Autonomous Weapon Systems – Legal, Ethical and Moral Challenges*, (VIJ Books (India) 2016).

[509] Dr U C Jha, *Killer Robots: Lethal Autonomous Weapon Systems – Legal, Ethical and Moral Challenges*, (VIJ Books (India) 2016).

[510] Samuel Gibbs, 'Elon Musk leads 116 experts calling for outright ban of killer robots' (The Guardian, Sun 20 Aug 2017) < https://www.theguardian.com/technology/2017/aug/20/elon-musk-killer-robots-experts-outright-ban-lethal-autonomous-weapons-war > accessed 20 August 2017.

[511] Samuel Gibbs, 'Elon Musk leads 116 experts calling for outright ban of killer robots' (The Guardian, Sun 20 Aug 2017) < https://www.theguardian.com/technology/2017/aug/20/elon-musk-killer-robots-experts-outright-ban-lethal-autonomous-weapons-war > accessed 20 August 2017.

Whilst AMs could make the battlefield safer, the flip side is that the technology could be abused and IHL infringed, with some experts fearing “that offensive weapons that operate on their own would lower the threshold of going to battle and result in greater loss of human life.”[512]

Currently military machines with combat capability are prohibited from being fully autonomous by design, ensuring there is always human input to make certain the laws of war, as per the Geneva Convention, are not breached, for example, by firing at targets within restricted fire zones.[513] However, today there are unconscious weapon systems already in use that have autonomy embedded in their ‘critical functions’ for selecting and attacking targets, although not activated.[514]

The International Committee of the Red Cross (ICRC) first highlighted concerns about unconscious autonomous weapon systems in its 2011 report[515], asking States to consider and emphasise fundamental legal, ethical and societal concerns presented by these weapon systems before they are developed and deployed. During March 2014, the ICRC organised an international expert meeting titled ‘Autonomous weapon systems: Technical, military, legal and humanitarian aspects’. It was attended by government experts from 21 States, with 13 individual experts who had a wide range of legal, technical, operational, and ethical expertise.

[512] Samuel Gibbs, ‘Elon Musk leads 116 experts calling for outright ban of killer robots’ (The Guardian, Sun 20 Aug 2017) < https://www.theguardian.com/technology/2017/aug/20/elon-musk-killer-robots-experts-outright-ban-lethal-autonomous-weapons-war > accessed 20 August 2017.

[513] Christopher McFadden, 'A Brief History of Military Robots Including Autonomous Systems' (Interesting Engineering, 06 Nov 2018) <https://interestingengineering.com/innovation/a-brief-history-of-military-robots-including-autonomous-systems> accessed 5 March 2019.

[514] ICRC, ‘Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects’ (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

[515] International Humanitarian Law and the challenges of contemporary armed conflicts.

The goal was to get a clearer understanding of the potential problems created by autonomous weapon systems and to share points of view.[516]

Despite not being conscious machines, it is the 'human out of the loop' category that is closet to MAMs and pose as much risk to complying with IHL as MAMs, do to accountability challenges when the technology takes action incompatible with IHL (e.g., used force that is not proportionate).

Within the UK, the definition of automated systems is that they follow pre-set rules with predictable outcomes; autonomous systems can understand environments, decide actions without direct human input, and exhibit unpredictable individual behaviours despite overall predictability.[517] The assessment of autonomy in weapons include factors such as the task performed, system complexity, and level of human oversight.[518] Automation of critical weapon functions has been present for years, although not necessarily linked to high complexity.[519] Autonomous weapon systems (AWS) are primarily used in limited, predictable contexts (e.g., defensive roles, attacking specific military installations), and targets are typically non-personnel, focusing on vehicles or objects.[520] Nevertheless, this level of autonomy does not include machine consciousness, so not subject to the TVAP, and requires

[516] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

[517] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

[518] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

[519] QinetiQ, 'MAARS weaponized robot' (QinetiQ, 2023) <https://www.qinetiq.com/en/capabilities/robotics-and-autonomy/maars-weaponized-robot > accessed 20 June 2023;
BAE Systems, 'Taranis' (BAE Systems, 2023) <https://www.baesystems.com/en/product/taranis > accessed 20 June 2023.

[520] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

a human to be in control and thus accountable under IHL. Consequently, it is the human deciding to use the autonomous machine, the action it should take, all with the full awareness of IHL, the values, and the implications.

4.11.1 Machine Consciousness, VAP and IHL

As a reminder, IHL is founded on the principles of distinction, proportionality, necessity and humanity. Machine consciousness, thus MAMs, present several challenges in relation to the application of IHL. This is both in relation to the original interpretation, which has humans at the core and requires MAMs to be aligned to our values, but also in respect to protecting MAMs, which was not the intention of IHL when originally drafted. This second facet is explored further in chapter 6.

With regards to IHL and the values it encompasses, MAMs will need to be designed and taught to align and comply with the principles:

- Distinction: MAMs must be capable of accurately distinguishing between combatants and non-combatants. This will require MAMs to understand multifaceted human behaviours, cultural contexts, along with the unpredictable nature of warfare, which already has been highlighted will be an extremely ambitious and laborious training activity, which it is argued here may prove prohibitive. There is also unlikely to be any margin for error allowed for MAMs.
- Proportionality: This principle requires a MAM to balance the potential harm to non-combatants against the military advantage gained, which demands a level of ethical reasoning, which MAMs will again need to learn.

- Necessity: There is a question over if MAMs would be able to be used against a State that could not afford to invest in the technology. On one hand, it could be argued that by using a MAM, with its advanced data processes, it could reduce the risk of harm and thus enables the waring State to achieve their legitimate purpose of conflict. However, it could be viewed as an excessive means and thus against the principles of IHL.
- Humanity: MAMs will have to be trained to recognise and understand the IHL compliant methods and means of warfare, in order to limit to limit suffering and injury. They will have to capable of evaluating situations where a combatant may wave a white flag, what constitutes the limiting suffering and injury, as well as obeying the Hague Convention of 1907. Suffering and injury may arguably be the hardest for a MAM to understand, as human suffering is unique due to psychological and biological factors, and hard to translate to MAMs, as highlighted is chapter 5.

However, the above has looked only at the requirement for MAMs to align to our values. If we truly want them to respect, uphold and see the value that we see, then we need to be an advocate for the benefits and proactively extend them to MAMs, which it is argued here is the true value alignment problem (TVAP). The TVAP is explored in chapter 6.

Other factors surrounding IHL are:

- Interpretation of Legal Principles: IHL can require complex interpretations of what constitutes proportionality, necessity, and distinction, which can be exacerbated the mist of battle and unfortunately can be misjudged, as in the case of Marine A. Human soldiers employ situational awareness, experience, and judgment to interpret the principles, and often have to make split-second decisions. MAMs will be required to have sophisticated interpretative capabilities, which not only consider the immediate situation and data, but also understands the broader ethical and pollical contexts. This is a requirement that is highly complex and uncertain.

- Accountability: Accountability is essential for maintaining compliance with IHL due times of war. The question of if a MAM commits a violation of IHL, who is held accountable, was discussed in chapter 3, and showed that MAMs will be able to meet the requirements of actus reus and mens rea, thus it would be unjust to hold the developers and/or the military operators responsible. Undeniably, the concept of MAM as moral agents (discussed in chapter 5), questions the currently established framework of accountability, as MAMs may take decisions beyond human foresight.
- Autonomy versus Control: Due to the implications of machine consciousness and thus autonomy, it is judged that the less control human operators will have over MAMs actions. It is this autonomy that tests the MAMs ability to ensure compliance with IHL, as their decisions may not align with human ethical standards and, further, focus on their own survival and preservation, as explored in chapter 5.
- Ethical and Moral Reasoning: Human compliance with IHL depends on ethical and moral reasoning, which is routed in human emotions, empathy, and recognition of human suffering. MAMs would need to replicate or be developed to genuinely possess an understanding of these moral values, in order to make decisions that comply with IHL principles. Yet, teaching such complex moral reasoning to MAMs is still viewed here as a fundamental challenge.

4.12 Summary

Humans have always driven the need for more sophisticated tools. The tools to date we used over hundreds of years have been inanimate, under our control and beholden to us. AM development is another step in our tool development; however, AMs have the ability to surpass our knowledge and capabilities and introduce new risks, as a result of their 'black box' characteristics.

Presently, computers are subservient, without consciousness, which we own and command, and with our rights clearly outweighing theirs. However, this may not always be the case, and we need to be prepared to accept this or simply stop the development of AMs now. We need to define the humane treatment of AMs and consider such questions as at what point might we consider deletion of AMs algorithms a form of mass murder? This is explored in chapter 5.

The chapter highlighted the complexity, rapid pace and trajectory of AM development and will test our 'extrinsic' controls and standards, including IHL and the values we hold dear. It is the view here that through our persistent development, AM/MAMs will develop consciousness, thus have awareness and be capable of self-reflection, and this will alter how we view AMs, their ethical status, our ability to use them, and brings to the fore the value alignment problem (VAP).

Whilst designing and developing machine consciousness will be a major feat, it is actually recognising the implications and consequences for our treatment of AMs and MAMs that is asserted here as the biggest challenge and where the author argues the true value alignment problem (TVAP) arises, especially with regards to IHL. Indeed, establishing consciousness initiates the argument for recognising personhood. This subsequently leads to realising agency and free will, which constitutes autonomy, and ultimately it is argued throughout this thesis, places an obligation to protect their 'life', principally under IHL for MAMs. This ethical facet of machine consciousness is discussed in chapter 5.

The research has shown that machine consciousness has not been specifically designed with IHL in mind, and indeed the 'black box' characteristics with AMs support this and likely to cause challenges in determining and understanding the degree of value alignment. However, it should be noted that nor is any human born to comply with IHL. Soldiers are educated and trained in IHL to understand and comply with it. Hence, it is argued here that both the MAM and human brain start off as a blank canvas, and the argument of nurture versus natured argued for both. Yet, humans typically choose to enlist in the armed forces and thus come to identify and uphold IHL, of which MAMs will not be afforded such choice. Therefore, there will need to be an extensive period of training to a yet to be defined and established MAM universal IHL standard to understand the nuances, and perhaps even a negotiation with MAMs as to why they should adhere to IHL. Thus, for MAMs to respect, align and uphold our IHL values, it is argued that we must demonstrate good conscience and equity and extend IHL and the inherent values, to them; This is argued here as the true value alignment problem (TVAP). Chapter 6 builds from this chapter and explores, and argues for, the TVAP for MAMs.

Despite the technology not being specifically designed with IHL compliance in mind, it is regarded as a requirement on the State to only deploy technology into the military environment that is deemed compliant; The State accepts the risk.

5. RA3: Examination into whether the current code of IHL ethics aligns with machine consciousness decision-making.

5.1 Introduction

Chapter 4 highlighted the significant shift from inanimate machines viewed as mere tools, to that of today's thinking machines, which are able to process vast amounts of complex data and decide on a course of action. Even with human oversight, ethical questions are already presenting themselves, thus, with each advancement taking us closer to a form of machine consciousness, far greater and more fundamental questions arise. Chapter 3, stated that, the law tends to derive from societal beliefs, norms, and values[521], which have an ethical root and encapsulates the value alignment problem (VAP). Consequently, this chapter aims to draw attention to the emerging ethical questions and the ethical challenges MAM decision making could present IHL.

Ethics is the "well-founded standards of right and wrong that prescribe what humans ought to do, usually in terms of rights, obligations, benefits to society, fairness, or specific virtues."[522] Viewing this definition in light of this research aim, it is seen here that ethics is a system of moral principles and values, that affect how we make decisions and lead our lives. Ethics has established foundations in the medical field, where it is accepted as comprising of the 4 core principles of beneficence, nonmaleficence, autonomy, and justice[523], of which

[521] P. Sales, 'CONSTITUTIONAL VALUES IN THE COMMON LAW OF OBLIGATIONS' (2024) The Cambridge Law Journal, 83(1) < https://www.cambridge.org/core/journals/cambridge-law-journal/article/constitutional-values-in-the-common-law-of-obligations/10695D32CEDAA3E2EEC391C212AF3925 > accessed 4 September 2024.

[522] Manuel Velasquez, Claire Andre, Thomas Shanks, S.J., and Michael J. Meyer, 'What is Ethics?' (Markkula Center for Applied Ethics, Santa Clara University, 1 January 2010) < https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/what-is-ethics/#:~:text=First%2C%20ethics%20refers%20to%20well,%2C%20fairness%2C%20or%20specific%20virtues > accessed 4 September 2024.

[523] B Varkey 'Principles of Clinical Ethics and Their Application to Practice' (2020) Med Princ Pract. 2021;30(1):17-28 < .https://pmc.ncbi.nlm.nih.gov/articles/PMC7923912/#:~:text=Beneficence%2C%20nonmaleficence%2C%20autonomy%2C

autonomy is of key interest within this research aim. The UN Ethics Office additionally include accountability and human rights to their ethical principles.[524]

As mentioned above, ethics is related to that which is good for individuals and society and is also considered as moral philosophy. From the wealth of research and opinions about ethics, trying to distil it into one sentence or phase appears provocative, however Majeed has make a very good attempt by saying that ethics can be “described as ‘the intrinsic control of good behaviour’. This is in contrast to ‘law’ that acts as the ‘extrinsic control of good behaviour’.”[525] Intrinsic control is at the core of how we view our world and our limitations. It is the fundamental constraints we place upon ourselves and thus what we expect from those around us. To understand intrinsic control, it is important to understand what it is to be human (e.g., personhood) and how we determine our freedoms and constraints (e.g., free will). As a result, decision making is first explored, looking at what is required for decision making and how decisions are made. This is followed by how the current code of IHL ethics aligns to MAM decision making.

5.2 Machine Consciousnesses and Decision-Making

%20and%20justice%20constitute%20the%204%20principles,the%20latter%202%20evolved%20later > accessed 4 September 2024.

[524] UN Ethics Office, ‘WHAT IS THE UN ETHICS OFFICE?’ (UN Ethics Office, 2024) < https://www.un.org/en/ethics/#:~:text=The%20UN%20Ethics%20Office%20promotes,and%20respect%20for%20human%20rights > accessed 4 September 2024.

[525] A.B.A Majeed, ‘Roboethics - Making Sense of Ethical Conundrums’ (2017) Procedia Computer Science, Volume 105, 2017 < https://doi.org/10.1016/j.procs.2017.01.227 > accessed 19 March 2019.

5.2.1 Ethics of Personhood, Moral Agency, and Consciousness

To be ethical, it is important in the first instance, that we understand the components that enable the human decision-making process (personhood and autonomy), before attributing characteristics and weight to AMs and MAMs, thus we start by exploring personhood.

The status of personhood grants individuals the highest level of moral value, of which is discussed further below. Young[526] assets that "personhood is a normative category in ethics and has a normative value… personhood is normative in that to acknowledge the other as a person is integral; it has value in seeing the other as a person,"[527] For humans to attribute status and rights to an AM/MAMs, along with respecting the decision-making of AMs/MAMs, establishing personhood is considered essential. Personhood encompasses several attributes, including consciousness, communication and self-consciousness.[528] Dennett[529] regards these as key conditions for moral personhood, although this may change with AMs, and possibly new moral status of 'robothood' could be founded. Kant viewed personhood as requiring moral agency, with a moral agent as a being who has a conscious understanding of right and wrong. Bryson et al[530] and Solaiman[531] are deeply engaged in the discussion around assigning

[526] G. Young, 'Personhood across disciplines: Applications to ethical theory and mental health ethics in Ethics, Medicine and Public Health' (2019) Ethics, Medicine and Public Health Volume 10, July–September < https://www.sciencedirect.com/topics/medicine-and-dentistry/personhood#:~:text=Personhood%20is%20a%20normative%20category,the%20person%20(after%20Kant) > accessed 4 September 2024.

[527]G. Young, 'Personhood across disciplines: Applications to ethical theory and mental health ethics in Ethics, Medicine and Public Health' (2019) Ethics, Medicine and Public Health Volume 10, July–September < https://www.sciencedirect.com/topics/medicine-and-dentistry/personhood#:~:text=Personhood%20is%20a%20normative%20category,the%20person%20(after%20Kant) > accessed 4 September 2024.

[528] D C Dennett, 'Conditions of personhood' [1976] The Identities of Persons, ed A. O. Rorty Berkeley, CA: University of California Press.

[529] 529 D C Dennett, 'Conditions of personhood' [1976] The Identities of Persons, ed A. O. Rorty Berkeley, CA: University of California Press.

[530] J J Bryson, M EDiamantis, and T.D Grant, T. D. 'Of, for, and by the people: the legal lacuna of synthetic persons' [2017].Artif. Intell. Law 25, 273–291. doi: 10.1007/s10506-017-9214-9.

[531] S.M Solaiman, 'Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy' (2017) Artif Intell Law 25, 155–179 (2017) <https://link.springer.com/content/pdf/10.1007/s10506-016-9192-3.pdf > accessed on 17th October 2018.

legal personhood to robots, both seeing some benefits and issues. Ultimately, Bryson et al see the difficulties in holding AMs accountable when they infringe the rights of humans as outweighing "the highly precarious moral interests that AI legal personhood might protect"[532], whilst Solaiman argues that "robots are ineligible to be persons, based on the requirements of personhood."[533]

Philosophers such as Kant view personhood as requiring moral agency. A moral agent is deemed as a being who has a conscious understanding of right and wrong. Ethics provides the principles and guidelines, which inform what is considered right or wrong, consequently guiding the moral agent in their decisions. Personhood and moral agency are separate from moral value. Humans have the right to life and are inherently valuable regardless of individual personality and character traits.[534] However, characteristics of moral value itself causes debate. Singer[535] asserts that moral value originates from the ability to feel pain or pleasure. Regan[536] considers that inherent moral value derives from being a conscious individual with a life that has importance to itself regardless of its usefulness to others, i.e., being the experiencing subject of a life. Whilst Warren[537] proclaims that sentient animals and foetuses have some inherent value due to meeting key qualities of her criteria. Warren argues that for an entity to be considered a person, thus have moral value, it must satisfy a number of the following qualities:[538]

[532] J J Bryson, M EDiamantis, and T.D Grant, T. D. 'Of, for, and by the people: the legal lacuna of synthetic persons' [2017].Artif. Intell. Law 25, 273–291. doi: 10.1007/s10506-017-9214-9.

[533] S.M Solaiman, 'Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy' (2017) Artif Intell Law 25, 155–179 (2017) <https://link.springer.com/content/pdf/10.1007/s10506-016-9192-3.pdf > accessed on 17th October 2018.

[534] John Harris, 'Wonderwoman and Superman,' (1992) Oxford.

[535] Peter Singer, *Practical Ethics, 2nd Ed*. Cambridge University Press (New York & Cambridge, U.K.: 1993).

[536] Tom Regan was an American philosopher who focused on animal rights theory.

[537] Mary Warren, 'On the Moral and Legal Status of Abortion' The Monist, 57 (1973).

[538] Megan-Jane Johnstone, *Bioethics: A Nursing Perspective* (7th Edition, Elsevier 2019).

A. Consciousness of events and objects, both internal and/or external, to the being and the ability to feel pain;

B. Reasoning – The ability to decipher moderately complex and new problems;

C. Self-motivated actions – Actions that are comparatively free from genetic or direct external control;

D. The ability to communicate via any means, "messages of an indefinite variety of types, that is, not just with an indefinite number of possible contents, but on indefinitely many possible topics;"[539]

E. "The presence of self-concepts, and self-awareness, either individual or racial, or both."[540]

Warren's qualities are of interest here as, Warren views A and B alone as necessary for personhood, although she does not assert that any of the qualities are unconditionally necessary. Nevertheless, she does insist that an entity who lacks all the above qualities is clearly not a person, thus hold little, if any, moral value. Warren's qualities are echoed, either in their entirety or individually, by other researchers and philosophers. Whilst quality B above is very easy to prove for AMs/MAMs, quality A, consciousness, along with humans acknowledging it, is their major hurdle to personhood, and consequently ethical treatment of them, as discussed in chapter 6.

[539] Megan-Jane Johnstone, *Bioethics: A Nursing Perspective* (7th Edition, Elsevier 2019).

[540] Megan-Jane Johnstone, *Bioethics: A Nursing Perspective* (7th Edition, Elsevier 2019).

For Kitwood[541], personhood is sacred and unique with every person having an ethical status. Kitwood argues that personhood ought to be treated with profound respect. At the core of Kitwood's theory is the moral concern for 'others,' which is essentially what this thesis is arguing is the 'true value alignment problem (TVAP)' with regards to AM/MAMs. Kitwood's theory is even more interesting as he focuses on dementia patents, thus those that ultimately do not have mental capacity, and it feels this focus can be transferred to AM/MAMs as they develop. Thus, in the first instance and to start getting humans comfortable, personhood can evolve and be recognised as AM/MAMs consciousness matures. Further, philosopher Peter French[542] believes a corporation can be a moral person, as it can be considered a moral agent, be morally responsible, and thus held accountable for its actions. This again can be used to support AM/MAM's bid for personhood, which helps uphold IHL values and closes legal gaps if MAMs are capable of being accountable and thus underpins the principle of ethics.

Harris has many interesting and somewhat controversial ideas regarding persons and non-persons.[543] He argues that a person who wants to live is wronged if killed, as they were divested of what they value. However, non-persons or "potential persons" as Harris affectionately terms unborn babies and babies, "cannot be wronged in this way because death does not deprive them of anything they can value, though this does not exhaust the wrong that might be done by infanticide."[544] Harris is gracious in at least considering those that would care for the baby may be wronged, but is steadfast in his view that the baby cannot be wronged with regard to infanticide; "If they cannot wish to live, they cannot have their

[541] T Kitwood, *Dementia Reconsidered: The Person Comes First* (Open University Press 1997).

[542] Peter French is an American philosopher.

[543] John Harris, *Wonderwoman and Superman* (1992) Oxford.

[544] John Harris, *Wonderwoman and Superman* (1992) Oxford.

wish frustrated by being killed."[545] Whilst we can agree that computers of yesterday and today may well 'not wish to live', for tomorrow's AM/MAMs we need to rethink how we classify persons and non-persons, and start opening our minds and conversations to the real possibility that artificial consciousness will emerge within most of our lifetimes, and that the concept of personhood will need to adapt and expand to include non-human beings. This is especially true for MAMs, where they will be deployed into situations where surviving is the aim, and where it is the view here that because of this, there is a greater ethical need to have their existence and values recognised and thus protected through IHL; The TVAP.

Harris's views on the beginning of life, as discussed in chapter 4, is of further interest here as when an AM/MAM becomes a 'being' is equally as challenging and controversial. Indeed, this will introduce a whole new level of complexity to AM ethics and the controls around how we create and destroy the controlling AM software before it even has a chance to be 'switched on' and start learning and functioning as an AM. Once again, and to get around potentially tying ourselves up in ethical tape, we could pursue Harris' opinion of rationalising the difference between abortion, infanticide and murder and applying his opinion that:

> "creatures other than persons can be wronged in other ways, by being caused gratuitous suffering for example, but not by being painlessly killed. This explains the difference between abortion, infanticide, and murder and allows us to account for how we benefit persons by saving the lives of the human potential persons they once were, but at the same time shows why we do not wrong the potential persons by ending that life, whether it be an unfertilised egg or a newborn infant."[546]

Therefore, as long as we destroy the newly created AM/MAM software without causing any pain or suffering to the AM/MAM software programme, (pain and suffering will need to be

[545] John Harris, *Wonderwoman and Superman* (1992) Oxford.
[546] John Harris, *Wonderwoman and Superman* (1992) Oxford.

defined for AM/MAMs and is discussed later), then we can regard ourselves as absolved of any ethical guilt or legal burden. In contrast, just as ending the life of a human is ethically and legally fraught, so potentially could ending the life of an AM/MAM that has come into existence. This will be discussed later.

5.2.2 The Ability to Suffer

Under the IHL principles of humanity, it is forbidden to inflict "suffering, injury or destruction not actually necessary for the accomplishment of legitimate military purposes"[547], which additionally is in the spirit of the nonmaleficence' core medical ethical principle, therefore MAM suffering must be minimised. Thus, it is argued here that IHL principles should be extended to MAMs and thus recognise and protect their value and 'life', which is considered by the author as the TVAP. As a consequence, careful consideration of what is deemed suffering for MAMs will need to be determined before military deployment. Extending IHL principles to accommodate and address what the author identifies as the TVAP is the focus of research aim 4 (chapter 6). The TVAP regards how we need to adapt our IHL values to consider them, their best interests, and welfare.

Applying Harris's[548] view on hybrids (Harris discusses a gorilla/human hybrid), he suggests that if the hybrid were to evolve into self-conscious beings, then they would equate to a non-human species of persons, but benefit from moral status, "the rights freedoms, protections,

[547] British Red Cross, 'International humanitarian law' (British Red Cross, 2024) < https://www.redcross.org.uk/about-us/what-we-do/protecting-people-in-armed-conflict/international-humanitarian-law#:~:text=Humanity%20forbids%20the%20infliction%20of,accomplishment%20of%20legitimate%20military%20purposes > accessed 4 September 2024.

[548] John Harris, *Wonderwoman and Superman* (1992) Oxford.

and obligations that persons possess in virtue of their personhood."[549] It is the view here that this helps signpost the way for AM/MAMs to acquire personhood. Harris makes a clear argument that:

> "The wrongness of harming pre-persons does not lie in the wrongness of harming potential persons but rather in the wrong of harming the actual persons that the pre-persons will become…The point is simply that it is not wrong to harm potential persons so long as the potential is never actualised."[550]

Again, this can be applied to AM/MAMs in that we will be creating 'pre-AM/MAMs' and we should consider the non-maleficence ethical principle in that we do not cause harm during their development, which endures once they switched on. Thus supporting the view in this thesis that we could have a duty to 'pre-AM/MAMs', as it is asserted here that they will become conscious AM/MAMs with development, so their potential should be respected and protected. It is agreed here that currently, suffering for humans can be considered worse than for animals and AM/MAMs, as humans always have a psychological dimension and that the very expectation of suffering causes suffering in itself.[551] This is because "persons suffer in anticipation of suffering, in memory of suffering, and because they are conscious of relative deprivation of freedom from suffering."[552] From an experiential perspective, Singer agrees that non-humans would suffer less; "The same experiments performed on nonhuman animals would cause less suffering since the animals would not have the anticipatory dread of being kidnapped and experimented upon."[553] If AMs will ever have a psychological dimension and suffer as a human would is not clear, but they may well have psychological interests.[554]

[549] John Harris, *Wonderwoman and Superman* (1992) Oxford.
[550] John Harris, *Wonderwoman and Superman* (1992) Oxford.
[551] John Harris, *Wonderwoman and Superman* (1992) Oxford;
Peter Singer, *Animal Liberation* (First published 1975, Bodley Head 2015).
[552] John Harris, *Wonderwoman and Superman* (1992) Oxford.
[553] Peter Singer, *Animal Liberation* (First published 1975, Bodley Head 2015).
[554] Discussed in section (b), Personhood and Rights.

5.2.3 Personhood and Rights

One of the crucial aspects of personhood is the rights and duties ascribed to it. Thus, should AM/MAMs be granted personhood, then it would be ethical that they too would enjoy the rights and be burdened by the duties, which means both AM/MAMs aligning to our values, in conjunction with humans extending our values and protections for them, which is again argued here as the TVAP. As mentioned previously, the TVAP relates to the need for us to adapt our values, specially IHL values for MAMs, to respect AM/MAMs autonomy, and consider their best interests and welfare.

Extending our values and protections is seen here as starting with recognising their existence, via personhood, after which we would have a duty consider and protect it, including under IHL for MAMs.

With the advancement of AM/MAMs, it is appropriate to reconsider their moral status and consider the point at which to assign personhood, especially if the pace of technology leads to a form of AM/MAM consciousness. To reiterate, personhood is the highest level of moral value as argued by Singer[555] and Regan[556]. Whilst Harris[557] and Singer[558] take a more controversial stance around moral value and who can experience suffering, Warren[559] takes an inclusive approach proclaiming that foetuses and sentient animals do have some inherent value. This opens the door for AMs/MAMs, as it shows that humans already accept the value in non-persons, so demonstrates a preparedness to see value beyond ourselves, which could

[555] Peter Singer, Practical Ethics, 2nd Ed. Cambridge University Press (New York & Cambridge, U.K.: 1993).

[556] Tom Regan was an American philosopher who focused on animal rights theory.

[557] J. Harris, 'Wonderwoman and Superman,' (1992) Oxford.

[558] Peter Singer, Practical Ethics, 2nd Ed. Cambridge University Press (New York & Cambridge, U.K.: 1993).

[559] Mary Warren, 'On the Moral and Legal Status of Abortion' The Monist, 57 (1973).

be carefully manoeuvred (or even manipulated) to reframe the value we see in AM/MAMs. Kitwood[560] shares Warren's view that every person has an ethical status. Moreover, Wooldridge[561] even argues AM/MAMs could be superior to humans, which is not a future Harris[562] and Singer[563] clearly imagine or hope for. Taking this further, French believes a corporation can be a moral person, as it can be considered a moral agent and thus held accountable for its actions, which is a step closer to acceptance of AM/MAMs[564] and easing some of the accountability tension under IHL. As accountability is key within both IHL and ethics, being able to hold MAMs accountable for their actions, again, goes a significant way in bridging the liability gap. Nevertheless, humans may feel uncomfortable with MAMs being exclusively accountable for their actions, which is discussed in chapter 3.

As mentioned above, Bryson et al[565] and Solaiman[566] are active supporters of assigning legal personhood to robots, although they both consider the practicalities of this akin to legal personhood much like an organisation. Further, some technologist and researchers, such as Coeckelbergh,[567] Darling,[568] and Gunkel,[569] argue that ascribing rights to AMs is positive. Indeed, Darling[570] claims that it aligns to our social values and treating AMs like our pets rather

[560] T Kitwood, Dementia Reconsidered: The Person Comes First (Open University Press 1997).

[561] Micheal Wooldridge, The Road to Conscious Machines: The Story of AI (Pelican Books, 2020).

[562] J. Harris, 'Wonderwoman and Superman,' (1992) Oxford.

[563] Peter Singer, Practical Ethics, 2nd Ed. Cambridge University Press (New York & Cambridge, U.K.: 1993).

[564] Peter French is an American philosopher.

[565] J J Bryson, M EDiamantis, and T.D Grant, T. D. 'Of, for, and by the people: the legal lacuna of synthetic persons' [2017].Artif. Intell. Law 25, 273–291. doi: 10.1007/s10506-017-9214-9.

[566] S.M Solaiman, 'Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy' (2017) Artif Intell Law 25, 155–179 (2017) <https://link.springer.com/content/pdf/10.1007/s10506-016-9192-3.pdf > accessed on 17th October 2018.

[567] Mark Coeckelbergh, *The Political Philosophy of AI: An Introduction* (Polity; 1st edition, 11 Feb. 2022).

[568] K Darling, 'Extending legal protection to social robots: the effects of anthropomorphism, empathy, and violent behavior towards robotic objects' (2012) We Robot Conference 2012, April 23, 2012, < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2044797 > accessed 10 November 2017.

[569] D J Gunkel, 'Robot Rights' [2018] MIT Press. doi: 10.7551/mitpress/11444.001.0001.

[570] K Darling, 'Extending legal protection to social robots: the effects of anthropomorphism, empathy, and violent behavior towards robotic objects' (2012) We Robot Conference 2012, April 23, 2012, < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2044797 > accessed 10 November 2017.

than as things is normal human behaviour, due to our tendency to anthropomorphise and ability to nurture.

Katz[571] supported by Smart and Richards[572], warns that how we relate to AMs is not consistent with how we relate to humans, with the social robot Sophia being an example, which is discussed later in this chapter. Pagallo's suggestion of a duty to nurture AM/MAMs, very much like a parent does for a child, rings true. It must surely be of benefit to all that we do nurture AM/MAMs and teach them to be 'good'[573], to benefit society[574] and to align to our values, but it is the view here that they must see a benefit to aligning to our values in order for them to do so. It would therefore become a positive obligation on every AM/MAM owner to do so, very much as a parent does for a child, which will create an obligation for us to teach them and train them on 'good' quality and valid data, and thus ethical. For example, not to train AM/MAMs on data that goes against British values, that is racist and incites hatred, however, what is considered 'good' data within the military context, where you are expecting MAMs to kill the enemy, is beyond the scope of this thesis, but remains a question that will need to be answered. Nevertheless, the ramifications of not correctly teaching AM/MAMs, could be as wide ranging as for a parent, e.g., court action and AM/MAMs taken away. Teaching AM/MAMs good behaviour will help instil the parameters of intrinsic control, which includes values, they must operate within. We need to consider the situation and outcome if an AM/MAM chooses to be 'bad' despite good teaching and how we would sanction this. As

[571] B. Katz, 'Why Saudi Arabia Giving A Robot Citizenship Is Firing People Up' (Smithsonian Magazine, 2 November 2017) < https://www.smithsonianmag.com/smart-news/saudi-arabia-gives-robot-citizenshipand-more-freedoms-human-women-180967007/ > accessed 2 November 2017.

[572] Neil M Richards and William D Smart, 'How Should the Law Think About Robots?' (10 May 2013) < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2263363> accessed on 17th November 2017.

[573] What is considered 'good' and 'bad' is actually quite subjective and influenced by so many factors, e.g., a person's upbringing, religion, values, morals, etc.

[574] Ugo Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts* (2013 edn, Springer.

mentioned previously, having personhood bestows right and duties, as a consequence, the will be accountable for the consequences of not upholding our values, just as humans do. This is the other side of the TVAP, but is not dissimilar to what we expect of humans in upholding our values. Nevertheless, criminal punishment is beyond the scope of this thesis. However, we will quickly need to get intrinsic control embedded and entrenched in their thinking and actions. This is particularly imperative to IHL and how we ensure MAMs make ethical decisions, with the values and principles of IHL at the fore, and which is dependent on 'good' quality teaching.

Singer's view is that not all humans are equal, thus we should stop pretending that they are. This consequently provokes conversations relating to ways we can make things more equal, ever mindful that we live in a world plainly devoid of equality. Indeed, IHL tries to level inequality in that there is a basic duty towards the welfare of the enemy. This conversation can then expand to include AM/MAMs and builds a case for giving them personhood by saying they are not equal to humans (yet!), but they do require consideration and protection from exploitation and harm. Indeed, developing AM/MAM equality in line with human equality, enables us to see a pathway forward for the protections offered under IHL and is ethical. Thus, not everyone will agree with MAMs being treated as equal to humans, but if the law states it is so, then people will just have to abide by it. Nevertheless, as mentioned, the law states that groups of people and individuals (e.g., women, transgender, etc.) are equal, yet that does not stop humans from not a treating them as such or ignoring the law. However, in both situations, it does not stop the person or AM/MAM, who is being treated as not equal, from acting ethically and being morally responsible. Indeed, it is the view here that the person not treating them as equal is unethical and ignoring their values, but that does not detract from their ethical status, value, or behaviour. Further, just because a small number of people, groups, or sectors, may be hesitant or even resistant to acknowledging AM/MAM personhood and rights, should not be a reason to slow down or halt the recognition of a entity's status and rights as this fuels the TVAP.

5.3 Autonomy and Agency

For Kant, autonomy is required for personhood, with autonomy being one of the key ethical principles. Chella and Manzotti's[575] work is influenced by Heisenberg[576], Laplace[577], and Spinoza[578]. They acknowledge Spinoza's view that people think they are free because they believe they are in control of their actions. Our belief of free will derives from our conviction that we can act otherwise from what can be expected.[579] Haladjian and Montemayor[580], and Signorelli[581], rightly warn that we should not assume that we will be able to control AM/MAMs (e.g., their thinking or actions), which highlights IHL and ethical accountability will be challenging.

[575] Antonio Chella and Riccardo Manzotti, 'Machine Free Will: Is free will a necessary ingredient of machine consciousness?' (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017.

[576] Antonio Chella and Riccardo Manzotti, 'Machine Free Will: Is free will a necessary ingredient of machine consciousness?' (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017.

[577] Pierre-Simon, Marquis de Laplace. (1749-1827), philosopher.

[578] B Spinoza, 'The Ethics' (1664/2009) Ethica Ordine Geometrica Demonstrata, New York, Dodo Press.

[579] Yoshiteru Ishida & Ryunosuke Chiba "Free Will and Turing Test with Multiple Agents: An Example of Chatbot Design" [2017] Procedia Computer Science. 112. 2506-2518.

[580] H H Haladjian, and C Montemayor, 'Artificial consciousness and the consciousness-attention dissociation' [2016] Conscious. Cogn. 45, 210–225. doi: 10.1016/j.concog.2016.08.011.

[581] Camilo Signorelli, 'Can Computers Become Conscious and Overcome Humans?' (2018) Hypothesis and Theory article Front. Robot. AI, Sec. Humanoid Robotics Volume 5 - 2018 < https://www.frontiersin.org/articles/10.3389/frobt.2018.00121/full > accessed 8 June 2019.

From the research, the common key elements for autonomy have been identified as those below and will be explored in detail:

- Agency:
    - Self-Determination – The ability to make one's own choices, free from controlling influences; Having sovereignty over oneself.
- Free Will
    - With free will comes moral responsibility, and refers to the human capability to exercise autonomy and make choices.

Both elements are analysed in the following subsections.

5.3.1 Agency

Human agency and ethics are complexly related, as both influence and inform human behaviour and decision making.[582] Agency is about the human ability to act autonomously and make free choices. It goes to the core of intrinsic control and how much we really have a choice over our decisions. It continues to be debated as to whether an individual acts as a free agent or acts in a manner prescribed by his social structure and is the structure versus agency debate.[583] This relationship between agency and structure is controversial in social

[582] Thomas M. Jones, 'Ethical Decision Making by Individuals in Organizations: An Issue-Contingent Model' (1991) The Academy of Management Review, vol. 16, no. 2, 1991, pp. 366–95. JSTOR < https://doi.org/10.2307/258867 > accessed 4 September 2024.

[583] Fred Dallmayr, 'Agency and Structure", [1982] University of Notre Dame Phil.Soc.Sci. 12 (1982) 427-438;
Alessandroa Morselli, 'The mutual Interdependence between Human Action and Social Structure in the Evolution of Capitalist Economy' (2014) Scientific and Academic Publishing p-ISSN: 2168-457X < article.sapub.org/10.5923.j.m2economics.20140201.02.html>;
Stephan Fuchs, 'Beyond Agency', [2001] University of Virginia.

science, where debates pivot around the fundamental and distinctive characteristics of an individual's action, e.g., "is human behaviour always intentional and conscious or are there cases of unconscious and repetitive actions? Are individuals independent making their own choices freely or is individual behaviour conditional?"[584] The agency-structure debate supports the argument for AM/MAM free will, as it focuses us to confront our own 'programming' through the structures and environments we grow and operate within. Thus, we will be forced to face up to the fact that free will is a mere illusion or even a myth. Hooker sees that an AM/MAM is autonomous, perceiving the world through various technologies (e.g., sensors) then acts, as with an agent, upon its environment and towards a specific goal[585], which is a view shared here.

According to Heidegger[586], humans are Dasein[587] in that it is the fact we exist or are simply in the world that defines us; 'Being-there' is central to Heidegger's philosophy. It relates to his theory of human agency and its significance in shaping our human lives.[588] Whilst Dasein can be a reasonable characteristic attributed to humans, it should be considered plausible that we may soon use the term to describe AM/MAMs. Although Owen and Owen make a strong argument for exclusively human agency and thus would not accept the TVAP, by saying, "no machinery or cyborg has the capability of formulating and acting upon decisions without human programming, and no cyborg qualifies as Dasein."[589] This is supported by Fuchs who views agency as requiring "consciousness, free will and reflexivity".[590] With the advancement

[584] Alessandro Morselli, 'The Mutual Interdependence between Human Action and Social Structure in the Evolution of the Capitalist Economy, Microeconomics and Macroeconomics' [2014] Vol. 2 No. 1, 2014, pp. 6-11. doi: 10.5923/j.m2economics.20140201.02.

[585] J Hooker, 'Autonomous Machines Are the Best Kind, Because They Are Ethical', (2016), Carnegie Mellon University < http://public.tepper.cmu.edu/jnh/agencyPost2.pdf > Accessed on 14 December 2017.

[586] Martin Heidegger (1889–1976), German philosopher.

[587] Dasein: "there-being". Humans are defined by the fact that they exist or are in the world and inhabit it.

[588] M Horrigan-Kelly, M Millar, & M Dowling, 'Understanding the Key Tenets of Heidegger's Philosophy for Interpretive Phenomenological Research' (2016) International Journal of Qualitative Methods, 15(1) < https://doi.org/10.1177/1609406916680634 > accessed 4 September 2024.

[589] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).

[590] Stephan Fuchs, 'Beyond Agency', [2001] University of Virginia.

of machine learning and AM/MAMs, it is very much the machine doing its own learning and decision making, thus not the programmers. Hooker believes this view is obsolete in the face of AM/MAMs with DL capabilities and considers an AM/MAM to be a free agent if we are able to justify and, "predict its behaviour on the basis of the controlling algorithms"[591], which is a viewed argued against here, as predicting behaviour based on algorithms seems to suggest a degree of covert human control.

Owen's[592] 'Genetic-Social' meta-theoretical framework provides a good and welcome challenge for MAMs and wider AMs, especially given Owen's view that no AM/MAMs will be Dasein. His framework is applied here to AM/MAMs and used to analyse the strengths, weaknesses and possibilities Owen's philosophical approach offers for recognising AM/MAM agency. Owen's framework derives from, "behavioural genetics, evolutionary psychology, the neuroscience of free will, and the philosophy of Heidegger in the task of 'building bridges' between the social and biological sciences."[593] The most prominent and applicable areas of Owen's framework for this thesis is his focus on biology and neuro-agency on human free will. Indeed, because his framework focuses on humans, he sets a high bar for measuring free will and therefore measuring an AM/MAM against his framework provides a sound benchmark for AM/MAM free will, highlighting areas that we will have to review when recognising AM/MAM free will and to make it ethical. This actually creates a framework for establishing when the TVAP initiates, as his framework can categorise AM/MAMs with free will and therefore the focus of the TVAP, and those that do not fall under the TVAP (e.g., unconscious

[591] J Hooker, 'Autonomous Machines Are the Best Kind, Because They Are Ethical', (2016), Carnegie Mellon University < http://public.tepper.cmu.edu/jnh/agencyPost2.pdf > Accessed on 14 December 2017.
[592] A leading international criminological theorist based at UCLan and part of my supervisory team.
[593] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).

car, laptop, or military aircraft). In contrast to Owen, Hirschi[594] concentrates on causality, looking at why people do not commit crimes and focuses on their constraints, such as status and socialisation.

Consequently, if Owen's framework can explain human agency, it could be revised to explain AM/MAM agency. Indeed, Owen's framework will be examined further to understand the theoretical reasoning, and applied to conceptualise AM/MAMs in understanding the meta-constructs, such as the biological variable (biological influence upon human behaviour that could be fed into AM/MAMs and how they are 'influenced'), psychobiography (asocial characteristics of a person – will asocial characteristics be evident in AM/MAMs?). Owen's term 'neuro-agency', which recognises influence of neurons upon human 'free-will'[595], is of special interest in this thesis due to highlighting the similarities between human and AM/MAM, and one of the key reasons his framework is exclusively focused on here. Further, Owen's concept of neuro-agency has important ethical implications, as it challenges conventional of free will, moral responsibility, and the dissimilarity between humans and machines. It is this recognition of the influence of neurons on human decision making, coupled with the emphasis on the similarities between human and machine consciousness, that sees neuro-agency bring to the fore ethical concerns. As a result, it appears a good starting point for illustrating AM decision-making processes. Indeed, despite Owen's reservations and his non-rectified view, his framework could be expanded to rationalise AM/MAM behaviour and decision making, which would aid understanding the TVAP and how we start aligning our values.

[594] Travis Hirschi, American Sociologist.

[595] Tim Owen, Crime, Genes, Neuroscience and Cyberspace (Palgrave Macmillan 2017).

5.3.2 Neuro-Agency

Owen makes a strong case for embracing a new metaconstruct, termed neuro-agency, in criminological analysis, which:

> "acknowledges research in the field of the neuroscience of free-will and posits a neural influence upon human decision-making. It is the contention here that an inherited impulsive disposition may predispose an actor to formulate and act upon potentially criminal decisions."[596]

Whilst Owen does stress 'human decision-making' and holds a strong position that biology is required for neuro-agency, the spirit of the new metaconstruct could be applied to potential future conscious AM/MAMs with ANNs and demonstrate that neuro-agency could be possible in conscious AM/MAMs. To reiterate, AMs, including MAMs, will be born from DL, so centred around ANNs, which originate from the complex structure and functions of our brain (discussed in chapter 4).

In explaining neuro-agency, Owen states that:

> "the metaconstruct, neuro-agency is employed in Genetic-Social metatheoretical reasoning as an acknowledgement of the neural influence upon human free-will. It is contended here that it is timely essential to acknowledge recent developments in the neuroscience of free-will and to abandon the 'old' term 'agency'."[597]

His notion of agency is a non-reified one, where actors or agents are distinguished as entities which can formulate and act upon decisions. Owen is by no means keeping neuro-agency sacred to humans, as he is very open to cetaceans and primates being incorporated if it can

[596] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).
[597] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).

be proved that they too are self-aware. Self-awareness is interlinked with ethics, as it influences decision-making and responsibility. Further, the author states that the TVAP holds that if AM/MAMs become self-aware, then we must reconsider how we treat them within ethical and legal frameworks.

5.3.2.1 'Simulated, non-human agency'

In Owen's recent publication[598], he introduces the concept of 'simulated, non-human agency, which refers to "programmed 'machine agency"[599], and acknowledges "thinking technology"[600] that, "employ a form of non-human, non-self-reflective, programmed agency."[601] However, Owen continues to hold firm that no machine can be Dasein and no machine can ponder its own finitude, which the author asserts will be unethical when considering AM/MAMs. Yet his concept of simulated, non-human agency fits neatly with Hallevy's first and second models and even complements Hallevy's third model[602], where the AM/MAM is fully legally responsible for its actions and omissions. However, this is a notion that Owen currently will not budge on and he reaffirms that machines cannot be held culpable for acts, as they were programmed by a human actor, and subsequently liability should be assigned to the human, thus here the intrinsic, ethical, control is that of the operators alone.

A question that arises from Owen's framework is, 'are decisions always made with neural influence or could some decisions just be made purely on evaluating the facts?' An AM/MAM could evaluate the facts via ANNs and make decisions and actually be more consistent and

[598] Tim Owen, and Jessica Marshall, *Rethinking Cybercrime. Critical Debates (*Palgrave Macmillan, 2021)
[599] Tim Owen, and Jessica Marshall, *Rethinking Cybercrime. Critical Debates (*Palgrave Macmillan, 2021)
[600] Tim Owen, and Jessica Marshall, *Rethinking Cybercrime. Critical Debates (*Palgrave Macmillan, 2021)
[601] Tim Owen, and Jessica Marshall, *Rethinking Cybercrime. Critical Debates (*Palgrave Macmillan, 2021)
[602] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

ethical than a human, who, through the very processes of living, has acquire conscious and sub-conscious biases and preferences. It is the view here that Owen's definition of neuro-agency could be expanded to include the ANNs influence upon AMs free will, if Owen could accept AM/MAM could be Dasein in the context of machine consciousness. Thus, his framework seems flexible enough to adapt to the inclusion of AM/MAMs and recognise their autonomy, which is viewed here as ethical and will align to IHL.

5.3.2.2 Agency and Decision Making

Owen has aligned free will with Dennett's[603] soft-compatibilist model of free will, which proclaims that free will and determinism can coexist without being incoherent. Thus, agents are morally responsible for their actions, as long as their actions are not a result of external coercion. AM/MAMs could be coerced by humans, just as humans are coerced by other humans, for example, by family, through religion, peers, etc. It is not unreasonable to expect AM/MAMs may feel a pressure or duty to behave/act in a certain way.

If Owen could see a future where ANNs could be a prime for free will, then his framework becomes even more powerful and future proof. However, Owen clearly states that the concept of neuro-agency is non-reified, stressing, "it avoids at all cost what Sibeon identifies as reification; the attribution of agency to an entity incapable of formulating and acting upon decisions," although Owen leaves the door slightly ajar for AM/MAMs by upholding that Sibeon's view that:

> "there are actually two main types of actors; individual human actors and social actors…such as organisations, families, committees, central government

[603] Daniel Clement Dennett III is an American philosopher.

> departments, professional associations and so on. As Sibeon (2004: 119) puts it, the decisions of such social actors 'shape much of the social, economic, and political terrain' in today's society. They are not mere aggregations of the decisions of individual human actors."[604]

Here Owen makes a distinction in what actions and decisions could be made, by emphasising, "the form of agency which social actors (such as a political party) are capable of, in terms of decisions and actions, is different to that of individual human actors and it is essential not to confuse the two."[605] Owen does seem to imply that biological neuro-agency need not be the limiting factor when looking at agency, but for his metaconstruct, he makes it clear that:

> "It cannot however, be regarded as an individual actor on par with a sentient human being capable of the kind of neuro-agency which human individuals require to formulate/act upon decisions. To imply or suggest so would be to engage in the 'cardinal sin' or reification[606]."[607]

It could be argued that some people make certain decisions based on the ethical principle of beneficence, so the idea of the 'greater good'. Thus, this is not what they as an individual would want, but for the best of the collective, hence demonstrating free will could have boundaries and can be supressed. Duress removes free will completely, along with any veto power. Physiological (e.g., stress) and psychological (e.g., emotions such as fear) play a vital role in exercising one's free will. An AM could perhaps reach the utopia of free will because it will not be shackled with such factors. For MAMs, they may feel little, if any, pressure from their commander or troop to carry out an action, despite not viewing the action as correct or in their best interest, which would have positive implications for IHL and in line the 'traditional' VAP. They can make a decision unshackled by human factors such as fear. Owen

[604] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).
[605] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).
[606] Reification forms part of Sibeon's framework and is, "the 'illicit attribution of agency to entities that are not actors or agents'".
[607] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).

asserts that, “the term neuro-agency is applied here to acknowledge the evidence for a neural influence upon free will/agency, but this position does not suggest that free will is an illusion.”[608] However, it is worth emphasising that free will does have limits. In reality, exercising ones free will can reduce another person’s free will, therefore our free will and consequently our autonomy, is influenced and controlled by our neurology, as just an AM/MAM will be controlled by its ANN. Being aware of, and accommodating, the influence and control of ANNs on MAMs is argued here as key facet in the TVAP and recognises their uniqueness and autonomy. Certainly, humans have different influences, which shape our thoughts and behaviours, yet regardless of this and as mentioned in chapter 1, there is an expectation that humans share certain universal values, which the United Nations stated as peace, freedom, social progress, equal rights, and human dignity.[609]

5.3.2.3 Hybrid Agents

We need new thinking, and this is shown through Brown lending her support to hybrid agents.[610] Although Owen is forthright in stating, “no cyborg or machine can ever be Dasein”[611], perhaps, as a first step, Owen would be open to viewing AM/MAMs as social actors. From this definition, they appear to be made up of a combination of both human and social actors, as both have a heavy influence.

Owen and Owen assert that:

[608] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).

[609] United Nations, ‘Universal values - peace, freedom, social progress, equal rights, human dignity - acutely needed, Secretary-General says at Tübingen University, Germany’ (2003) United Nations Press Release < https://www.un.org/press/en/2003/sgsm9076.doc.htm > accessed 3 September 2024.

[610] Sheila Brown, ‘The criminology of hybrids: Rethinking crime and law in technosocial networks’ (2006) Theoretical Criminology, 10(2), 223–244 < https://doi.org/10.1177/1362480606063140> accessed on 17 November 2017.

[611] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).

> “no computer can be described as a “who” that is shaped and formed by existence in time, a creature with a past, its being accessed by means of an existential analytic, rather than a “what”, like some material object in space’.”[612]

This view is correct for today’s machines but lacks validity for the AM/MAMs on the horizon of tomorrow. Further, Owen contends that “material and non-human agency which attributes causal powers to physical objects (Pickering 2001) such as stones must be repudiated.”[613] Whilst one can accept that a stone, with no ability to make a decision or act upon a decision, could be repudiated, it seems ill-fitting to categorise an AM as an object. AM/MAMs will be expected to operate alongside the humans, including in the military, and make decisions based on their interpretation of the situation, just as a human soldier will be expected to. In reality, huge reliance and responsibility will be placed on MAMs to make IHL related judgement decisions, so we should be careful how we categorise them. Indeed, would we want to have people, regardless of status, killed by mere ‘objects’ during conflict, yet burden them with the weight of complex judgements, which is argued here as not aligned to IHL values? The two views seem at odds. It is argued here that we are all programmed to some degree by our environment and social factors, which is therefore little different to AM/MAMs.

5.3.2.4 ‘Rethinking’

With machine consciousness, it will no longer be the human that programmes the machine; indeed, they will programme and teach themselves, formulating and acting upon their own decisions. As a result, if we want AM/MAMs to align to our values, predominately IHL, then it

[612] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).
[613] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).

is argued here that we need to show them the benefits and extend IHL to MAMs. Here Brown is more forward thinking and declares that:

> “Transformational interfacing technologies (cybernetics, genetic engineering, digital visualization, satellite communications, convergent mobile communications, virtual environments) demand a rethinking of our heavily policed criminological boundaries. We need to dissolve the ‘scientific’ theories and the ‘social’ theories in order to grasp where we are now; and that is immutably in the technosocial. Above all, this is a world where the ‘objects’ and the ‘subjects’, the ‘social’ and ‘scientific’, of criminology’s purview are co-extensive and symmetrically active.”[614]

“Rethinking” is vital to ensuring future just wars and to the ethical treatment of all taking part in conflict. Brown’s alternative view to the concept of agency is viewed by Owen as being:

> “limited because it does not acknowledge a neural influence upon free-will and additionally Brown’s concept of agency is reified. Arguably, there has not occurred the ‘merging’ between human actors and machinery in cyberspace that Brown imagines.”[615]

Yet her ideas are interesting for AM/MAMs, with her views on the merging of humans removing some of the initial obstacles that drown the discussion. Brown advocates the idea of actors and actants, where actants, “are simultaneously informational and organic entities in the deepest sense, technology co-extensive with the human sensorium or self/personhood.”[616] Brown illustrates her view with Stone’s thought-provoking example of physicist Stephen Hawking, in saying that Hawking’s medical condition renders him reliant on his portable computer system, thus Brown asks where is Hawking? as she ponders that " a

[614] Sheila Brown, ‘The criminology of hybrids: Rethinking crime and law in technosocial networks’ (2006) Theoretical Criminology, 10(2), 223–244 < https://doi.org/10.1177/1362480606063140> accessed on 17 November 2017.

[615] Tim Owen, Crime, Genes, Neuroscience and Cyberspace (Palgrave Macmillan 2017)

[616] Sheila Brown, ‘The criminology of hybrids: Rethinking crime and law in technosocial networks’ (2006) Theoretical Criminology, 10(2), 223–244 < https://doi.org/10.1177/1362480606063140> accessed on 17 November 2017.

serious part of Hawking extends into the box on his lap . . . where does he stop? Where are his edges?'"[617]

Brown looks to remove the human 'body' from the equation, which is positive for AM/MAMs. Indeed, the actant-body is a leap forward for AM/MAMs, and Brown tries to see-off Owen's doubts by arguing that:

> "One obvious objection to granting equivalence to things and people is that things are not sentient; however one looks at it, it is the essentially human monad that experiences pain, fear and suffering, not the pixel or the machine. Consequently, the interface between people and things in the essential matters of ethics is particularly problematic, for how is the 'essentially human' to be isolated?"[618]

As previously emphasised, AM/MAMs may not experience pain and suffering as we understand or interpret it, but that does not mean that they are unable to suffer. For example, suffering could take the form of reduced processing capacity due to being denied a software upgrade. This type of suffering (e.g., reduced processing power) could be 'real' for the AM/MAM, even if not perceived as suffering to us. Understanding and acknowledging this is argued here as a vital aspect of the TVAP, so they do not suffer needlessly.

It is this actant-body that lights the way for personhood and opens us up to the possibility that machine consciousness should be valued and protected. However, nowhere in Brown's 2006 article does she touch on free will and consciousness, which is tackled and argued

[617] Sheila Brown, 'The criminology of hybrids: Rethinking crime and law in technosocial networks' (2006) Theoretical Criminology, 10(2), 223–244 < https://doi.org/10.1177/1362480606063140> accessed on 17 November 2017.

[618] Sheila Brown, 'The criminology of hybrids: Rethinking crime and law in technosocial networks' (2006) Theoretical Criminology, 10(2), 223–244 < https://doi.org/10.1177/1362480606063140> accessed on 17 November 2017.

against by Owen. Brown goes as far as saying that humans and non-humans should not be separated, but does not confront machine autonomy, merely that production of knowledge affords agency. Owen highlights the thinking behind criminal activity by saying:

> "when the actor, in this case perhaps a potential criminal offender in cyberspace, is faced with a decision, 'a consideration-generator whose output is to some degree undetermined produces a series of considerations, some of which may of course be immediately rejected as irrelevant by the agent [consciously or unconsciously]' (Dennett 1981: 295). The considerations that are selected by the potential criminal offender as, 'having a more than negligible bearing on the decision' then play a significant part in a reasoning process, 'and if the agent is in the main reasonable, those considerations ultimately serve as predictors and explications of the agent's final decision'."[619]

Although this thought process could easily be followed by MAMs for deciding whether to take action or not. Indeed, MAM decisions and actions "will have more than a negligible bearing", as MAMs will be making decisions based on their own information. It is also argued that AM/MAMs could be reflective agents as Owen asserts that human are, stating:

> "human beings are reflexive agents with the neuro-agency to choose not to engage in criminal activities where they believe that the rewards are outweighed by negative outcomes or actions offend moral prohibitions. Agency, in turn, is certainly influenced by inherited constitutional variables."[620]

AM/MAMs will be influenced by their environmental 'experiences' and inherited constitutional variables from their programmer as a human does from their parents/guardians. Owen's recipe for agency includes human psychology and physiology. Owen sees humans as "reflexive agents with the agency to choose not to engage" [621] in undesirable behaviour. Owen views agency as "influenced by inherited constitutional

[619] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).
[620] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).
[621] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).

variables… include neural influences"[622] , with the 'mind' being an outcome of the connections between "the human brain, the human body, and the environment."[623]

Owen's blocker for machine consciousness and free will seems to be neural influences and therefore not aligned the authors views. Nevertheless, if Owen could park this human physiological aspect and accept ANN, then his framework is suddenly elevated as applicable to AM/MAMs and in doing so, would not risk ignoring or infringing their personhood, which would also help address the author's TVAP. It is argued here that AM/MAMs will be able to reflect, and arguably be better at recalling facts, as they will not view things from a place of emotion or bias, thus not distorting events. This factor could prove vital and essential within the military context and significantly reduce the human factors (e.g. emotions) that lead to Marine A situations.[624]

Owen argues that Brown focuses on examples that include artificial technology, but that there "is no case where the technology has not been programmed by a human actor" [625]. Owen accepts that in a very narrow sense some advanced technology may be able to 'think', he holds fast in stating that no technology is capable of being Dasein, thus "contemplating its own finitude as being in time."[626]

[622] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).
[623] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).
[624] Discussed in the Chapter 3, Part II
[625] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).
[626] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).

5.3.2.5 The Biological Variable

With human decision making, biology plays an influencing factor. Indeed, the biological variable within Owen's framework relates to "the evidence from evolutionary psychology and behavioural genetics for a, at least in part, biological basis for some human behaviour."[627]

There is no biological influence for AM/MAMs, nevertheless, AM/MAMs will be influenced by their physical factors such as memory and processing capacity, software version, and hardware limitations, along with their environment (a form of AM neuroplasticity!). When thinking and applying genes to AM/MAMs, we could consider their DNA to be their software code and algorithms, which would influence their behaviour, especially if you look at the values, objectives and background of the original programmer. Owen also sates that:

> "it is essential to recognise that, 'genes do influence behaviour'… a role for the environment (part of the 'complexity' and 'richness' referred to by Wilkie) which is 'massively important' (Ridley, ibid) in the sense that genes are 'switched on' by social cues."[628]

An AM/MAMs code/algorithm could also be 'switched on' by cues/triggers from its environment or interaction with human or other AM/MAMs, which could then change its behaviour. This could have both positive and negative effects when looking at IHL. For example, a MAM deployed in a specific military situation, that has dormant training data stored in its 'brain', but which is switched on when faced with a specific situation. This could be deemed unethical if it is deliberately hidden from the MAM and to be triggered at a given

[627] International Journal of Criminology and Sociological Theory, 'Towards a New Sociology of Genetics and Human Identity' [2013] Vol. 6, No.3, June 2013, 68-80 68

[628] International Journal of Criminology and Sociological Theory, 'Towards a New Sociology of Genetics and Human Identity' [2013] Vol. 6, No.3, June 2013, 68-80 68

time. For example, The MAM is unaware it is programmed to behave in a certain way when triggered. It is the view here that this would go against the MAMs autonomy and free will and argued here as akin to brainwashing[629] and part of the TVAP that needs to be addressed.

As highted in chapter 2, Owen also views that psychological factors, such as childhood traumas, will not factor into AMs/MAMs in the same way as they influence humans, although they will learn from their experiences, both good and bad. AMs/MAMs will learn from all experience and look towards a better outcome should they meet situations again. They will not suffer psychological pain or trauma in the same way as a human, but, again, does not discount they may suffer in ways we would not perceive to be suffering, which is part of the TVAP. This feeds into the 'Humanity' principle of IHL and the non-maleficence principle of ethics, which subsequently feeds into the TVAP.

5.3.3 Free Will, VAP and TVAP

This section explores free will with the aim of understanding wider views and the relationship with the VAP and what the authors terms as the true value alignment problem (TVAP).

Free will refers to the human capability to exercise our autonomy and make choices, which are truly our own. With free will comes moral responsibility, and thus ownership of our good and bad actions and decisions, and where Christian[630] sees the VAP evolving from. Fisher's[631]

[629] Marcia Holmes, 'Hiding in Plain Sight' (*BBK*, 29 June 2015) < http://www7.bbk.ac.uk/hiddenpersuaders/blog/hiding-plain-sight/ > accessed 4 September 2024.

[630] Brian Christian, *The Alignment Problem: How Can Artificial Intelligence Learn Human Values?* (September 2021, Atlantic Books)

[631] John Fisher, 'Free Will and Moral Responsibility' (*UCL*, 2024) < https://www.ucl.ac.uk/~uctytho/dfwFischer2.html > accessed 4 September 2024.

assertation that "if we do not have free will, then there is no such thing as moral responsibility…without free will there is no moral responsibility,"[632] and subsequently no ethics. This is agreed with here and supports the author's view of free will. Fisher goes on to state that "if moral responsibility exists, someone has free will. Therefore, if no one has free will, moral responsibility does not exist."[633] As discussed, responsibility and accountability are key foundational aspects of ethics, the VAP and thus attributes AM/MAMs should possess. AM/MAMs possessing these attributes is argued here as laying the foundations for being recognised as having autonomy and being moral agents, which is asserted here as ethical and further both address the VAP but manoeuvres the conversation and focus to the TVAP, which is further explored in chapter 6.

As previously stated, it could be considered that humans are programmed to align with values by their parents, peers, society, religion, etc., therefore the difference between programming a machine and a human is reduced when looking at free will; both make decisions based upon the knowledge they have gained and the environment they are in.

An AM's existence is another problem for Owen's framework, which Brown does not confront in her research. AM/MAMs will physically exist, but Owen raises the question of birth and death, saying that "what it means for a human actor to be is to exist temporally in the time between birth and death."[634] As previously mentioned, 'birth' could be the first time the AM/MAM is activated (e.g., built and switched on). Death is probably harder to identify as the software/algorithms will most likely merge and replicate into other AM/MAMs. Perhaps

[632] John Fisher, 'Free Will and Moral Responsibility' (*UCL,* 2024) < https://www.ucl.ac.uk/~uctytho/dfwFischer2.html > accessed 4 September 2024.

[633] John Fisher, 'Free Will and Moral Responsibility' (*UCL,* 2024) < https://www.ucl.ac.uk/~uctytho/dfwFischer2.html > accessed 4 September 2024.

[634] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).

AM/MAMs will be better at understanding their finitude than us, as they will require humans to give them access to power, data, etc. An MAM that is destroyed in conflict could be said to ‘die’, however it is likely that data could be retrieved from it (e.g., either from the hardware itself or from the cloud). Thus, an MAM could be revived, yet this will depend upon human intervention and the willingness to undertake such action. This serves to highlight the complexities around artificial ‘life’.

Owen stressed the complexities of the human life within his framework, expressing that:

> “If an embryo is unable to formulate and act upon a decision, under the definition employed here [Owen’s framework] it is clearly reification to regard the entity as an actor…It seems difficult to regard a foetus as an actor in the sense employed in the framework, that is to say, an entity that is, in principle, capable of formulating and acting upon decisions.”[635]

Strictly speaking, an embryo cannot be Dasein when applying Owen’s view, and therefore aligns more with Harris’s views of embryos and babies. An AM/MAM would not be as limited or restricted as an embryo and would be able to formulate and act upon decisions. Indeed, no one will expect an embryo to be burdened with the time critical and life and death decisions of armed conflict that a MAM could be faced with. As a result of this, it is the view here that it is even more important we ensure the ethical treatment of MAMs for the start of any development and in doing so, they will learn to recognise and align to our values from conception.

Although Owen claims that cyborgs/machinery are not Dasein, he does not deny that some advanced technology can 'think' and produce new potential beyond programming by human

[635] International Journal of Criminology and Sociological Theory, ‘Towards a New Sociology of Genetics and Human Identity’ [2013] Vol. 6, No.3, June 2013, 68-80 68.

actors. Nevertheless, Owen maintains that unless the AM/MAM has the type of self-aware consciousness that allows consideration of its own finitude, reflexivity and the possibility of integration within a 'moral community', then Owen cannot see them being regarded as Dasein. However, Owen's new term 'non-human simulated actors', recognises the emerging ability of AM/MAMs to think beyond their programming. Nevertheless, he is firm in his view that without consciousness of the Dasein type, no technology can be held culpable for its behaviour. Thus, with the aim of accommodating the technology advances, Owen generously offers a distinction between the agency of human beings (e.g., Dasein, conscious, self-aware 'beings in time') and the 'simulated' agency of AM/MAMs, which as is argued throughout, again does not align to the author's view. Yet, it must be stressed that the author recognises that human agency is inherently subjective, contextual, and dynamic, but where the author differs to Owen, is that the author maintains that AM/MAM agency could be equally so. Owen views simulated agency as instrumental, functional, and lacking the existential factors of human agency, therefore bolding holding that the agency of human beings and simulated agency of machines do not align. Despite Owen respecting that machines (AM/MAMs) can imitate functional aspects of agency (for example, decision-making, goal-orientated behaviour), he again holds strong that they lack the ontological depth of human beings (self-aware, morally responsible, and time-bounded)[636], which influences our views on agency and moral status, and the interrelationships of humans. As a result, the then shapes how we consider ethical questions and challenges.

[636] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).

Nonetheless, Owen’s current concept of 'Neuro-Agency' and ‘Simulated, Non-Human Agency’ is a good starting point to distinguishing between the agency of human beings and the 'simulated' agency of AM/MAMs. His Neuro-Agency acknowledges the neural influence upon human free will (more like 'Free Won't'...) and the parameters of intrinsic control, yet discounts AM/MAMs from having this neural influence. If it can be proved that certain Cetaceans and Primates are self-aware too, one wonders whether Owen would be willing to incorporate them under his term of neuro-agency. It is hoped here that this opens the door for AM/MAMs and is the start of recognising their free will, moral responsibility, and thus lays the groundwork for discussing the TVAP.

Decision making includes choices and options that are freely chosen. We are all too quick to think we have free will and are autonomous, but do humans really have autonomy? Is our intrinsic control not evolved from our culture, values, religion, upbringing, laws, environment, relationships, finances, etc.? Is autonomy and free will a mere illusion and in reality, we are not able to do as we choose? We may well think we are making free choices for ourselves, but perhaps we are conditioned to make those choices and they are in fact deterministic. There are positive aspects of autonomy that we must maintain to be deemed as having autonomy and thus by default must be free from the negative aspects. Indeed, The Centre for Professional Ethics at UCLan supportively state:

> “Personal autonomy is, at minimum, self-rule that is free from both controlling interference by others and from limitations, such as inadequate understanding, that prevent meaningful choice. The autonomous individual acts freely in accordance with a self-chosen plan, analogous to the way an independent government manages its territories and sets its policies. A person of diminished autonomy, by contrast, is in some respect controlled by others or incapable of deliberating or acting on the basis

> of his or her desires and plans. For example, prisoners and mentally retarded individuals often have diminished autonomy.[637]

Our belief of free will derives from our conviction that we can act otherwise from what can be expected.[638] Shida and Chiba endorse an interesting hypothesis that humans couple consciousness with free will to circumvent any impasse or any recurring actions within a short timeframe.

Ishida and Chiba assert:

> “Free will, which is a significant aspect of consciousness, plays an important role in self-related problem solving. Free will allows a problem-solving agent to choose options freely without any constraints… One can feel free will when one can react or can think in an unexpected manner. Thus, we believe current machines may not have free will because their interactions occur in a limited capacity that may be deterministically defined.”[639]

Nevertheless, the author argues that this lack of free will would not amount to ethical practice for the AM/MAMs, which are the focus of this thesis.

Ishida and Chiba judge consciousness and free will as characteristics to allow problem-solving not attributable to the current day AM/MAMs, which seems a fair verdict. They attribute the argument on whether machines are able to display or utilise free will to the argument about whether humans can generate truly random numbers: “Once we have defined free will, it

[637] J Varelius ‘The value of autonomy in medical ethics’ [2006] Med Health Care Philos. 2006;9(3):377-88. doi: 10.1007/s11019-006-9000-z. Epub 2006 Oct 11. PMID: 17033883; PMCID: PMC2780686.

[638] Yoshiteru Ishida, Ryunosuke Chiba, 'Free Will and Turing Test with Multiple Agents: An Example of Chatbot Design' [2017] Procedia Computer Science, Volume 112, 2017, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2017.08.190 > accessed 3 March 2019.

[639] Yoshiteru Ishida, Ryunosuke Chiba, 'Free Will and Turing Test with Multiple Agents: An Example of Chatbot Design' [2017] Procedia Computer Science, Volume 112, 2017, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2017.08.190 > accessed 3 March 2019.

would be difficult to admit that the machine behaviour based on the definition indicates a free will."[640] This statement causes a certain amount of concern and nervousness, as surely if an AM/MAM does meet the definition, then we must recognise it has free will or we could end up in a speciesist position, risking discrimination and ignoring welfare for the belief of our own uniqueness. This is further viewed here as unethical and underpins the author's assertation of the TVAP. We should not impart our own ideas of what human free will is on non-humans, as we will overlook the uniqueness of non-humans, in addition to under respecting and valuing them, as demonstrated in Goodall's ground-breaking chimpanzee work.[641] We need to measure ideas of free will and consciousness by looking at what those characteristics look like for that animal, AM/MAM, etc, and not try to universally apply our human centric rules and metrics.

While reflecting on machine free will, Ishida and Chiba[642] favour the following definition that, "any approximate system that can simulate the behaviour of the target system, but can behave otherwise, against the simulated behaviour."[643] They see this definition as presenting the free will questions as a "nondeterministic problem for which only probabilistic statements are possible. It implies that finite state machines cannot simulate a system with free will."[644] This view of free will as a 'system's' ability to deviate from deterministic behaviour, leads to

[640] Yoshiteru Ishida, Ryunosuke Chiba, 'Free Will and Turing Test with Multiple Agents: An Example of Chatbot Design' [2017] Procedia Computer Science, Volume 112, 2017, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2017.08.190 > accessed 3 March 2019.

[641] Dame Jane Goodall is primatologist who revolutionised our understanding of chimpanzees.

[642] Yoshiteru Ishida, Ryunosuke Chiba, 'Free Will and Turing Test with Multiple Agents: An Example of Chatbot Design' [2017] Procedia Computer Science, Volume 112, 2017, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2017.08.190 > accessed 3 March 2019.

[643] Yoshiteru Ishida, Ryunosuke Chiba, 'Free Will and Turing Test with Multiple Agents: An Example of Chatbot Design' [2017] Procedia Computer Science, Volume 112, 2017, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2017.08.190 > accessed 3 March 2019.

[644] Yoshiteru Ishida, Ryunosuke Chiba, 'Free Will and Turing Test with Multiple Agents: An Example of Chatbot Design' [2017] Procedia Computer Science, Volume 112, 2017, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2017.08.190 > accessed 3 March 2019.

ethical questions regarding the responsibility and accountability of AM/MAMs if they can simulate free will, which this thesis focuses on.

The AM/MAMs' ability to use technology (e.g., sensors) to make a choice of the action to take, is a clear indication of agency.[645] This may not be comfortable for all, but perhaps Owen and Owen will need to review their statement that, "no cyborg qualifies as Dasein" in light of the emerging technology. Even Fuchs acknowledges that his idea that agency requires "consciousness, free will and reflexivity"[646] is open to challenge, and argued here as unethical, with Hooker quashing such ideas in the new world of AM/MAMs. This challenge to the requirements of agency could result in a too wide interpretation with more things having agency, and thus ethical status, than was originally intended or foreseen, e.g., random number generators.

Anderson and Anderson's[647] acceptance that "general ethical principles" along with periodic AM updates and training, can minimise ethical dilemmas, seems a workable and acceptable solution given scientific knowledge to date. This is an advancement on humans, who you cannot 'upgrade' themselves and override out of date beliefs, vales, and behaviours; we tend to cling to beliefs whether they are justified or not (e.g., homophobia). Vitally, Savirimuthu cautions about neglecting issues of diversity and gender when developing AM/MAMs.[648] Certainly, we do not want to transfer our bias to AM/MAMs and thus must be mindful of this throughout their development.

[645] J Hooker, 'Autonomous Machines Are the Best Kind, Because They Are Ethical', (2016), Carnegie Mellon University < http://public.tepper.cmu.edu/jnh/agencyPost2.pdf > Accessed on 14 December 2017.

[646] Stephan Fuchs, 'Beyond Agency', [2001] University of Virginia.

[647] Susan Anderson and Michael Anderson, 'The Consequences for Human Beings of Creating Ethical Robots,' (2007) aaai.org < https://www.aaai.org/Papers/Workshops/2007/WS-07-07/WS07-07-001.pdf> accessed on 20 October 2017.

648 Savirimuthu, Joseph, "Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence" Patrick Lin, Keith Abney and Ryan Jenkins (eds), International Journal of Law and Information Technology, Volume 26, Issue 4, Winter 2018, Pages 337–346, <https://doi.org/10.1093/ijlit/eay011> accessed 9 January 2023.

Hooker's explanation that "an agent is a being that is capable of action, and action is the exercise of agency" is encouraging as it allows for AM/MAMs to be included in the definition and thus be treated ethically. Nevertheless, Brozek and Jakubiec[649] do not see AM/MAMs as being creators of their own actions and, further, do not view AM/MAMs as being capable of having emotions and having the facility to feel pleasure and pain. On the other hand, using our scales for measurement of AM/MAMs is ignoring the biological limitations of people. Not having biological limitations could be seen as a significant advantage for AM/MAMs, thus is dangerous, unethical, and detrimental towards AM/MAMs to use human-centric scales. In fact, Brozek and Jakubiec steer well clear of addressing the problems their theory has when considering people with behavioural or psychological issues, e.g., Asperger syndrome. Indeed, people with Asperger syndrome may well identify or even bond with an AM due to not having emotional preconceptions or baggage, although this may not be considered ethical.

Luck and d'Inverno's[650] idea that a program is an autonomous agent if they can operate independently of the program users and can modify their goals according to changes in circumstance, is free will in action and is the anticipation of many technologists for future AM/MAMs. Luck and d'Inverno are forward thinking and welcoming of the technological advances of AM/MAMs, which needs to be replicated in us all, and is considered here a commendable view. We must accept that other entities are currently able to act with free will (e.g., a random number generator) even though it is not on the scale we view ourselves as operating on.

[649] Bartosz Broozek and Marek Jakubiec, 'On the legal responsibility of autonomous machines' [2017] Artif Intell Law (2017) 25:293-304.

[650] Michael Luck and Mark d'Inverno, 'A Formal Framework for Agency and Autonomy' (1995) aaai.org < www.aaai.org/Papers/ICMAS/1995/ICMAS95-034.pdf> accessed on 20 October 2017.

Building on Chella and Manzotti's idea that self-determination is vital to free will, along with the ability to fulfil own goals, is thought-provoking especially when they question our control over our behaviour, stressing that our behaviour is influenced by many causes. We accept that our parents, peers, environment, religion etc., influence and mould our intrinsic, ethical, control system, thus our influences steer us towards a choice, so can we really say we have free will? We could consider that our parents have 'programmed' us to behave in a certain way and hold certain values, very much like a software programmer who programmes an AM with rules via code. Further, Savirimuthu draws attention to how our agency and autonomy is heavily influenced and even curtailed through our use of technology and the algorithms underpinning them. Indeed, we are influenced and act upon the information gathered and processed by technology, which is then subsequently fed back to us as options or an action to take, e.g., health tracking app, Facebook advertisements[651].

This is akin to AM/MAMs, who too will have their agency and autonomy influenced via the data their sensors gather. This 'influence' could be seen as us being controlled by our technology, in that it is eroding our free will and agency, which could be considered unethical. Nevertheless, there are those individuals who have enacted their free will and managed to break free from these moulds, yet it is arguable that from leaving one set of moulds, they ultimately end up in another, new, mould. For example, if a person left a religious group and decided to become an atheist, it is arguable that they have simply moved into a new mould and will believe and act within that new moulds boundaries, such as not celebrating the religious aspect of Easter. There is so much overt and covert structure, rules, expectations,

[651] Savirimuthu, Joseph, Do Algorithms Dream of 'Data' Without Bodies? (January 25, 2017) < https://ssrn.com/abstract=2905885 > Accessed 9 January 2023.

etc., that we humans experience throughout our lives, that it is very hard to whole-heartly say we act and operate completely with free will. Interestingly, much of this overt and covert pressure will not burden AM/MAMs, perhaps making them 'freer' and consequently more aligned to ethics. AM/MAMs lack of human emotions such as shame, worry and embarrassment, could enable them to make better decision, which are free from burden.

We appear very much wedded to the idea that humans, and only humans, are conscious and have free will. We need to move on from binding and categorising things according to our view of the world, scales and metrics, and instead look to develop new tools and methods to measure that are open and allow for inclusiveness (e.g. AM/MAMs), or does our fear stop us from doing this? Perhaps we take comfort in the fact that at present we are the only entity that can meet the scales we have created, giving us an air of superiority and allowing us to behave entitled. Regardless of our motivation, it is held here that we need to develop and implement new tools and scales at pace and in readiness for MAMs, as it is the view here that MAMs will face the greatest limitations on their autonomy and free will, along with the greatest risks to their 'life'.

Interestingly, and as has been suggested earlier, a random number generator could meet the definition for free will, if it could output numbers otherwise than expected. It is the ability to do as we choose and act with unpredictability that is a key competent of free will.[652] Ishida and Chiba claim that chatbots, due to being built as an open system that interfaces with the internet, are systems that can behave as if they have free will. Thus, the agents (chatbots) can

[652] Yoshiteru Ishida, Ryunosuke Chiba, 'Free Will and Turing Test with Multiple Agents: An Example of Chatbot Design' [2017] Procedia Computer Science, Volume 112, 2017, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2017.08.190 > accessed 3 March 2019.

expand their world models based on the internet, which creates more options that can be selected. It is here where we can truly see free will in operation and the presence of 'unpredictability' that is normally (and hopefully) absent in most humans. Indeed, unpredictability is not a trait we often celebrate or look for in humans, and is a term frequently used to discuss those with severe mental health conditions such as schizophrenia[653]. Nevertheless, for a random number generator, unpredictability is exactly the trait you want, along the ability to be truly free to generate a number. Yet for a MAMs, where they are working in unpredictable situations, IHL brings in a degree of predictability to conflict via rules and principles, thus one would want MAMs to act within the IHL principles and ethics when faced with the unpredictable on the battlefield, e.g., limiting harm to non-combatants.

Farnsworth's barrier to AM/MAMs having free will is the absence of being a Kantian whole.[654] Yet, just as we do with children, we can teach AM/MAMs which acts are right or wrong, and further teach them a duty to act appropriately and make good decisions. For example, for MAMs to make decisions within the IHL guardrails, such as treatment of captured enemies, irrespective of the good or bad resulting consequences. We should not see this as disbarring them from free will, as previously stated, we 'program' our children, so it is argued here that programming AM/MAMs is no different. However, it is accepted here that children are allowed development time, which affords them the ability to make mistakes and learn from them, yet as raised before, AM/MAMs may not be afforded the luxury of mistakes, and there

[653] C Matthia, Herbert Angermeyer Matschinger, 'The effect of violent attacks by schizophrenic persons on the attitude of the public towards the mentally ill' (1996) Social Science & Medicine, Volume 43, Issue 12 < https://doi.org/10.1016/S0277-9536(96)00065-2 > accessed 17 November 2017.

[654] Keith Farnsworth, 'Can a Robot Have Free Will?' (2017) Entropy MDPI < https://pure.qub.ac.uk/portal/files/130713217/entropy_19_00237_v3.pdf > accessed on 1 November 2017.

could be an expectation that they have the required level of decision making (to be determined) from the moment they are switched on. With this in mind, it is argued here that MAMs should not be deployed until they have had some degree of assessment to test their decision making and that they act appropriately. Between different cultures and religions, we argue what is appropriate behaviour and values (e.g., circumcision, arranged marriage) and, agreeing with Hooker, Kantian language implies ideas that are irrelevant to this particular argument, specifically, that some things should never be done, regardless of the good consequences they produce.

Chella and Manzotti[655], and Farnsworth[656] agree that consciousness is not necessarily linked to free will, which is an area discussed at length within the field of medical ethics and found that ethical treatment of patient should continue regardless of the state of consciousness.[657] Farnsworth gives support to AM free will by regarding that consciousness does not appear to be a requirement. When we look at consciousness and free will in humans, we often find an underlying battle, for example, a person may be conscious, however they may not have free will, so they may be forced to do things against their will (e.g., slavery, abuse). It seems hugely naive to think that all humans, just because of the mere fact that they are born as humans, instantly have and display free will. This is not true and in fact many live and die having never once exercised, let alone maximised, their free will. In addition, some people do not have the mental capacity to understand free will, or could have it taken away from them, under the

[655] Antonio Chella and Riccardo Manzotti, ‘Machine Free Will: Is free will a necessary ingredient of machine consciousness?’ (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017.

[656] Keith Farnsworth, ‘Can a Robot Have Free Will?’ (2017) Entropy MDPI < https://pure.qub.ac.uk/portal/files/130713217/entropy_19_00237_v3.pdf > accessed on 1 November 2017

[657] M Farisco, C Pennartz, J Annen, et al, ‘Indicators and criteria of consciousness: ethical implications for the care of behaviourally unresponsive patients’ (2022) BMC Med Ethics 23, 30 < https://doi.org/10.1186/s12910-022-00770-3 > accessed 4 September 2024.

ethical principle of beneficence as encapsulated under s2 and 3 of the Mental Health Act 1983.

Hooker[658], Farnsworth[659], and Basl[660] see free will as currently exclusive to humans but do see the potential for it to be extended to AM/MAMs as the technology advances. Perhaps in some ways AM/MAMs will exercise more freedom than we do, as they would not be burdened with emotions such as guilt, love or prejudices, thus would not make decisions based on them. Perhaps they will in fact teach us what free will actually is and how our life could benefit and be freer. With AM/MAMs having the benefit of open learning and the ability to expand their world view via the internet, they could be more liberal and offer support to groups of people who feel isolated and rebuffed by society. This very much echoes Anderson and Anderson[661] who view the development of machine ethics as helping us to understand our identity and the meaning of behaving ethically, leading to advancement of research into ethics theory.[662] This will challenge us and our values, but that can only be a good thing and reduce our complacency and speciesism. It will make us recognise that we are not the only entity to have free will and make decisions, and that there is still so much more learning to do about our existence and what it is to be human.

[658] J Hooker, 'Autonomous Machines Are the Best Kind, Because They Are Ethical', (2016), Carnegie Mellon University < http://public.tepper.cmu.edu/jnh/agencyPost2.pdf > Accessed on 14 December 2017.

[659] Keith Farnsworth, 'Can a Robot Have Free Will?' pg1, (2017) Entropy MDPI < https://pure.qub.ac.uk/portal/files/130713217/entropy_19_00237_v3.pdf > accessed on 1 November 2017.

[660] J. Basl, 'Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines,' (2014) Journal of Philosophy and Technology, 27(1), 79-96. 2014 < https://philpapers.org/archive/BASMAM.pdf > accessed on 10th December 2018.

[661] Susan Anderson and Michael Anderson, 'The Consequences for Human Beings of Creating Ethical Robots,' (2007) aaai.org < https://www.aaai.org/Papers/Workshops/2007/WS-07-07/WS07-07-001.pdf> accessed on 20 October 2017.

662 Anderson Susan and Anderson Michael, 'The Consequences for Human Beings of Creating Ethical Robots,' (2007) aaai.org < https://www.aaai.org/Papers/Workshops/2007/WS-07-07/WS07-07-001.pdf> accessed on 20 October 2017.

Eidenmuller believes that 'smart robots' are not only able to take purposive actions, but they can "exhibit 'moral agency': they seem to understand the consequences of their behaviour, and they have a choice of actions"[663]. This view is encouraging and once again inclusive and accepting of AM/MAMs. Children do not often see the consequences of their behaviour, yet we assume that they will grow to understand this. Therefore, it seems necessary to allow AM/MAMs to grow and develop like infants and not immediately measure them by the standards of an Oxford University Professor from day one of their 'life'. As discussed earlier in this chapter, for MAMs, they should not be deployed until they have had some degree of assessment to test their decision making and ensure they act appropriately. Eidenmuller's acknowledgement that acceptance of robot legal personality is "based on a utilitarian conception of 'the good' or whether it rather is based on a humanitarian/Kantian vision according to which not everything that is utility-maximizing is necessarily the better policy,"[664] continues the Kantian whole debate, but we should see AM/MAMs as one day having legal personality and be accepting of that day, celebrating that we have in fact created another being through our own vision, although this does place a burden of responsibility to nurture upon us. It could be argued that by us nurturing AM/MAMs, we take away their free will, but it is the view here that our responsibility to nurture is only for their initial development and interactions with their environment, as they make sense of their world.

[663] Horst Eidenmüller, 'Robots' Legal Responsibility' (University of Oxford, 08 Mar 2017) < https://www.law.ox.ac.uk/business-law-blog/blog/2017/03/robots%E2%80%99-legal-personality > accessed on 1 November 2017.

[664] Horst Eidenmüller, 'Robots' Legal Responsibility' (University of Oxford, 08 Mar 2017) < https://www.law.ox.ac.uk/business-law-blog/blog/2017/03/robots%E2%80%99-legal-personality > accessed on 1 November 2017.

5.4 IHL Ethical Implications of MAM Decision-Making

Personhood, moral agency and autonomy are fundamental to the discussion of MAM consciousness. Currently, IHL assumes decision makers exert moral judgment and empathy, which are qualities that MAMs will find challenging. Indeed, it is the ability to comprehend the human cost of war, coupled with acting with compassion, that is crucial in ensuring the protection of the values underpinning IHL during conflict. It has been argued by Umbrello[665] that If MAMs are entrusted with life-and-death decisions, then a risk arises that ethical consequences may be overlooked in preference to efficiency and strategic advantage. However, if we wish MAMs to align to our IHL values, then we need to demonstrate the benefits and respect their autonomy, thus extend and align IHL values to protect MAMs; the TVAP. This is heightened, as MAMs will be designed to be deployed in warfare unlike a human solider, so their autonomy limited by our need for protection against hostile States. The alignment of IHL ethics with MAM consciousness decision making presents several challenges:

5.4.1 Moral Agency

As discussed earlier in this chapter, AMs and MAMs presently lack moral agency, and consequently, they do not possess consciousness, emotions, or empathy. It is these human qualities that are at the core of ethical decision making, which relies on understanding and valuing human life and dignity. However, this chapter also argued strongly that AMs and MAMs will eventually meet the requirements for personhood and subsequently be a moral agent, therefore recognising and assigning moral agency will be fair and equitable. As a result

[665] Steven Umbrello, 'No Machine Should Choose: Defending Human Dignity in the Age of Autonomous Weapons' (Worldonfire.org, 19 September 2024) < https://www.wordonfire.org/articles/no-machine-should-choose-defending-human-dignity-in-the-age-of-autonomous-weapons/ > accessed 23 January 2025.

of becoming a moral agent, a MAM could be wronged under IHL, if the ethical considerations of IHL do not take into account the life and value of MAMs; the TVAP.

5.4.2 Accountability

Currently, IHL does not address accountability in the context of MAMs (a human would be accountable for unconscious MAMs), thus theoretically undermining the principles of justice. From the research, it is highlighted that keeping MAMs under control and aligned to our IHL values require nurturing by humans and understanding of the benefits of the values from MAMs. A State will hope that they cannot simply decide to do something of their own 'free will', or exercise their "free won't", thus it is argued here that we will need to be able to predict their behaviour in the battlefield more so than human soldiers, due to the higher standard they will likely be held to. This will place a significant burden on the training of MAMs. Nevertheless, MAMs could in fact teach us to be more humane and their lack of emotions could result in better decisions and action when in conflict. Indeed, Sassòli states:

> "A robot cannot hate, cannot fear, cannot be hungry or tired and has no survival instinct. The robot can delay the use of force until the last, most appropriate moment, when it has been established that the target and the attack are legitimate. Robots do not rape. They can sense more information simultaneously and process it faster than a human being can."[666]

Sassòli's view illustrates the benefits of MAMs, including the ethical benefits of use, which are in human favour, yet it is the view here that conscious MAMs will have a survival instinct and wish to uphold its autonomy. In fact, despite not being considered living or conscious, some

[666] Marco Sassóli, 'Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified' (2014) International Law Studies, US Naval college, INT'L L. STUD. 308 (2014) < https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=1017&context=ils > accessed 10 November 2017.

computer viruses are programmed to have a self-preservation instinct, which demonstrates a degree of autonomy, for example a computer worm such as the infamous Stuxnet.[667] This has benefits and drawbacks; Benefits for the person creating the virus in that the virus will replicate and penetrate systems to meet its goals, but drawbacks in that the pace it can permeate and cause significant harm, if designed to, e.g., the WannaCry ransomware attack on the NHS.[668]

Data processing will prove a significant advantage for MAMs, as human soldiers will become too overwhelmed with all the information and the decisions they will be expected to make. Nevertheless, the States developing and deploying MAMs must take precautions to avoid the enemy interfering with the MAM and ultimately using it against them and their citizens. There would be an obligation to stop an unlawful attack by an MAM, just as there would be to stop a rouge soldier. However, this could prove more difficult due to the advanced technology they will be equipped with, along with their data processing ability, which could enable them to outwit their commander and/or troop. Expanding upon this, a MAM could malfunction or act on something it has unintendedly learnt. Here the question over liability and punishment lies and highlights a potential accountability vacuum. Heyns states that:

> "This will not be acceptable because it will mean that the underlying values – the protection of humanitarian values and the rights to life and dignity – are in effect rendered without protection. War crimes are not crimes if there cannot be prosecution."[669]

[667] Stuxnet is a computer worm that was originally aimed at Iran's nuclear facilities and has since mutated and spread to other industrial and energy-producing facilities. https://www.mcafee.com/enterprise/en-gb/security-awareness/ransomware/what-is-stuxnet.html#:~:text=Stuxnet%20was%20a%20multi%2Dpart,and%20monitoring%20electro%2Dmechanical%20equipment.

[668] NHS England, 'NHS England business continuity management toolkit case study: WannaCry attack' (NHS England, 21 April 2023) < https://www.england.nhs.uk/long-read/case-study-wannacry-attack/ > accessed 4 September 2024.

[669] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

Thus, if it can be argued that if such an accountability vacuum is a basic factor for the use of MAMs, then this is seen here as unethical and will provide further reasoning to keep them as illegal weapons. Whether this is realistic in modern day warfare, is yet to be answered, however, being 'one war behind' is going to be a constant nagging pressure. Further, it is easy to imagine a future where people blame the MAM for a course of action and do not see a human as accountable. Indeed, Heyns asserts that;

> "The argument would be that the machine took its own decisions, which are unpredictable, not because computers act randomly, but because the environments in which they operate are so complex that all possible interactions between the system and the surroundings cannot be foreseen."[670]

Whilst this argument of unpredictability may apply to human decision-makers that operate within very complex situations and pressurised contexts, it is again the view here that MAMs will be held to a higher standard and therefore error tolerances will be close to zero, if not zero, which is argued here as not in alignment with the ethos of ethics. This zero tolerance is in line with the fully autonomous cars (non-conscious) and the current expectations of the public, but seen here as too simplistic a view for MAMs. Should things go wrong during authorised, regular usage of an MAM, liability will be called into question, as discussed in detail in chapter 3.

Whether we should safeguard against deferring to MAMs where value judgments are required is yet to be determined, although, it could be argued that their lack of emotion could lead to fairer decisions being made. Nevertheless, safeguarding for both our and the MAMs sake, specifically the use and decisions of MAMs outside of specified boundaries and outside

[670] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

the limits of their intended purpose, is responsible. In fact, evidence from the use of drones has shown that, “unmanned systems can easily be deployed in areas without a nexus to armed conflict, while the more permissive targeting rules of IHL are invoked to justify their use in ways that are impermissible in terms of IHRL.[671]

The "black box" nature of MAMs will complicate the understanding of how their decisions are made. As discussed in chapter 4, the lack of transparency raises ethical concerns and questions around the level of trust humans to have in MAMs decision making processes, especially when decisions have a life-and-death outcome.

5.4.3 Distinction

When deciding to take action, MAMs must be capable of accurately distinguishing between combatants and non-combatants, and between civilian and military objects. How MAMs will decide between the categories is presently unclear. Being able to tell if a person is acting under duress or questioning their action, is a nuance humans pick up on through instinct and feelings, but this will be a significant challenge to teach a MAM this. With regards to protecting MAMs, it should be reiterated here that IHL;

> “protects those who do not take part in the fighting… It also protects those who have ceased to take part, such as wounded, shipwrecked and sick combatants, and prisoners of war.

[671] ICRC, ‘Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects’ (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

> These categories of person are entitled to respect for their lives and for their physical and mental integrity. They also enjoy legal guarantees. They must be protected and treated humanely in all circumstances, with no adverse distinction."[672]

Thus, it is argued here that to be ethical, a damaged MAM or a MAM that has been taken prisoner, should also be afforded protected. As such, a State will have a duty to respect the integrity of the MAM, which could include access to power, not to dismantle the MAM or try to access its data, and very much aligns to all 4 of the core ethical principles.

5.4.4 Proportionality

When contemplating the action to take, MAMs will need to decide upon the most appropriate means and methods of warfare to employ, which must not be disproportionate to the military advantage sought.[673] This again is a judgement-based decision, dependant on the situation and context, and therefore not black and white, which will result in a training challenge for MAMs. Furthermore, the use of MAMs could intensify existing inequalities and vulnerabilities, for example, not as technically advanced States may suffer disproportionately from the consequences of MAM centred warfare and the strategic advantage it presents, which could be raise ethical challenges around the justice and fairness in the application of IHL.

[672] ICRC Advisory Service, 'What is International Humanitarian Law?' (2004) ICRC < https://www.icrc.org/sites/default/files/document/file_list/what-is-ihl-factsheet.pdf > accessed 4 September 2024.
[673] ICRC, 'Proportionality' (ICRC.org, 2024) < https://casebook.icrc.org/a_to_z/glossary/proportionality > accessed 4 September 2024.

Extending the proportionality principle for the protection of MAMs, the means and method used to disable or disarm enemy MAMs must not be excessive, for example, switching off a MAM compared to blowing it up or undertaking a form of cyber warfare.

5.4.5 Necessity

MAMs will need to understand and be able to balance, at what point they may resort to the means and methods that are necessary to achieve the legitimate aims of the armed conflict.[674] This principle also shares many similarities with the ethical principle of beneficence.

These decisions may arguably be easier for MAMs to make, as they will be equipped with technology (e.g. sensors, radar) that will enable them to analyse the conflict and run multiple scenarios quicker and better than humans. However, this may result in action been taken quicker than under a human weighing up the decision and outcome, resulting in greater harm and thus unethical.

When applying IHL for the protection MAMs, protection could adapted to include limiting activities such as hacking or realising a computer virus, which again very much aligns to all 4 of the core ethical principles.

[674] ICRC. 'Military Necessity' (ICRC.org, 2024) < https://casebook.icrc.org/a_to_z/glossary/military necessity#:~:text=The%20“principle%20of%20military%20necessity,prohibited%20by%20international%20humanitarian%20law > accessed 4 September 2024.

5.4.6 Humanity

Currently IHL assumes that decision makers employ moral judgment, draw from the values, and show empathy, of which empathy is likely to be the most challenging for MAMs and the hardest to teach them. MAMs will need to comprehend and consider of the human cost of war and make decisions that reflect compassion and limit suffering, to ensure the protection of IHL principles during armed conflict. There is a risk that should MAMs be entrusted with life-and-death decisions, then ethical considerations may be overlooked in preference of military efficiency and strategic advantage.

Humanity is argued here as the key principle to extend to MAMs, which will demonstrate our respect for their 'life', and is at the heart of the TVAP. When exploring humanity in regards to MAMs and limiting their suffering, we must also take a macro view alongside the micro view (e.g., individual characteristics), as distress for an MAM could be related to corrupt code or interrupted power. Looking at the environment MAMs will operates in, along with their technological needs (e.g., access to power) will enable a framework and context to view the suffering and distress.[675] As stressed, MAMs will not choose to be MAMs, so if is the belief here that upholding and respecting their 'life' and autonomy is therefore of highest importance.

5.5 Summary

[675] Peter Singer, Animal Liberation (First published 1975, Bodley Head 2015).

This chapter explored intrinsic control, its elements and the unique ethical issues surrounding AM/MAMs. It drew from the well-established ethical principles from medical ethics, which has been built upon to cater for other general ethical situations or expanded, as per the United Nations ethical principles.

Understanding how we form our intrinsic control is complex, so imparting this into AM/MAMs will be highly challenging and risky. Nevertheless, it is paramount we do this if AMs are to respect humans, our values, and value their own existence. It is argued here that AM/MAMs will evolve to have personhood and a form of consciousness. As a result, it is further asserted that we should be open to this thinking and plan now for this eventuality.

As a consequence, it was asserted here that AMs will have personhood and the right to autonomy when AMs meet the required elements, as this is viewed here as ethical, just and humane. We need to be mindful of their development and not ignore their rights purely because it serves us. The chapter highlighted that Kant believed that autonomy is required for personhood. The status of personhood grants individuals the highest level of moral value and that the overriding view is that consciousness, reasoning, the ability to communicate and self-awareness are vital for personhood. Nevertheless, whether and when personhood is granted is still reliant on humans recognising and respecting this, which is challenging given humans find it difficult to often recognise and respect the personhood of other humans, for example, the recent Russian conflict and their treatment of Ukrainians.

The chapter has highlighted that traditionally autonomy comprises of agency and free will, contained within a biological entity. It is this biological element that appears to be a major stumbling block for many. Whilst AM/MAMs will not be autonomous from a biological and physical viewpoint, in that they cannot re-construct themselves or replicate themselves (e.g., create a new physical AM/MAM), they will be capable to do this in the non-physical sense (e.g., can replicate their own software and repair their faulty code) and could consider their DNA to be their software code and algorithms. AM/MAMs will be influenced and controlled by their physical factors such as memory and processing capacity, software version, and hardware limitations, along with their environment. AM/MAMs will learn from their experiences, both good and bad, which is why showcasing the benefits of aligning to our values is vital.

Complex, but arguable less emotive, is recognising the 'birth' of AMs and if we can destroy the controlling software before it has a chance to be switched on. It is the view here that as long as we destroy the newly created AM software without causing any pain or suffering to the AM software programme, then we can regard ourselves as absolved of any guilt or legal burden.

The chapter showed how human suffering is viewed as worse than animal suffering and therefore by default AM/MAM suffering. This is because humans always have a psychological dimension. If AM/MAMs will ever have a psychological dimension and suffer as a human would is not clear, but they may well have psychological interests. In determining non-human suffering, we must also take a macro view alongside the micro view (e.g., individual characteristics), as distress for an AM/MAM could be related to factors such as corrupt code or interrupted power. Indeed, AM/MAMs may not experience pain and suffering as we understand or interpret it, but that does not mean that they are unable to suffer and this suffering would be ‘real’ for the AM. Consequently, it is argued here that this will have significant implications for IHL, especially the principle of humanity.

We are fairly confident in assuming today’s AMs do not have consciousness and feel pain, but this may be a reality in the near future and failing to recognise this may create a situation where we unknowingly torture, neglect and abuse AM/MAMs, for example, this could result from not properly maintaining AMs, allowing them to overheat or fatigue. Thus, the evolution of AM/MAMs needs to be strictly monitored and controlled and ensure that systems are put in place to limit abuse of power. Indeed, it is the view here that failing to do so would undermine the IHL values and principles, creating injustice, which goes against the objective of IHL to reduce inequalities; This is the TVAP.

This chapter found that agency is about the human ability to act autonomously and make free choices. Owen’s framework was explored to appraise AM/MAM agency, despite Owen’s views that no machine will ever be Dasein.[676] His framework sets a high bar for arguing AM/MAM agency but one that this chapter has demonstrated can be comfortably reached. Owen’s definition of neuro-agency could be expanded to include the ANNs influence upon AM/MAMs free will. Owen’s new concept of simulated, non-human agency fits neatly with Hallevy’s first and second models and even complements Hallevy’s third model, where the AM is fully legally responsible for its actions and omissions.[677]

Further, Owen’s biological variables and behavioural genetics could not only explain free will and neuro-agency for humans, but, if slightly adapted to incorporate variables and traits of

[676] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).

[677] Discussed in detail in Chapter 4: Prof. Gabriel Hallevy, ‘Virtual Criminal Responsibility’ (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

AM/MAMs, go a long way to explain free will and agency in AM/MAMs. Indeed, anti-reductionists believe that not all attributes of a system can be explained in terms of its parts and interactions.[678] This insinuates there can be something bigger at play and this could be free will for AM/MAMs.

Should Owen decide to revise his belief that a machine cannot ever be Dasein, then his framework could be adapted for AM/MAMs and have greater reach. Certainly, coupled with Hallevy's models, this would provide a comprehensive and powerful framework for understanding AM/MAM free will and assessing accountability and resultant legal liability.

Our intrinsic control is tied to our perception of free will. Free will is a much-debated concept in humans and in reality, some humans may never exercise any free will over their lives, which could lead to controversy in recognising AM/MAM free will. However, we appear very much wedded to the idea that humans, and only humans, are conscious and have free will. We need to readjust our thinking and move away from binding and categorising things according to our view of the world, scales and metrics, instead looking to develop new tools and methods to measure, which are open and allow for inclusiveness of AM/MAMs. Indeed, some of the ways AMs are described in this thesis may now be considered obsolete or deemed inappropriate for future AM/MAMs. As AM/MAMs develop, we must look for and recognise their free will or we could end up in a speciesist position, risking discrimination and ignoring welfare for the belief of our own uniqueness, which further undermines the sentiment of IHL. We should not impart our own ideas of what human free will is on non-humans, as we will overlook the uniqueness of non-humans. AM/MAMs will form and exercise their own intrinsic

[678] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).

control, but that will be learnt from us, thus we shoulder a huge responsibility for teaching AM/MAMs, our values and why they should align to them. The justification for MAMs aligning to our values, needs to incorporate and convey a mutual benefit, so they feel invested and bought in to upholding our values.

Basl stated that when we have created artificial consciousness that has capacities similar to ours and that encounters the world very much like us, then we should accept that consciousness as a moral patient.[679] It is not just human beings with consciousness that are moral patients, as animals are moral patients, and we are obliged to consider their interests, ensuring minimal harm.

With AM/MAMs making decisions that will have a profound impact on our lives, e.g., medical AMs, it seems only feasible that they should be afforded rights at a minimum of those of a corporation. Yet future AMs will have a form of consciousness, so just as with animals, it appears just they have further rights.

If MAMs will ever be truly free to act or not act is questionable. It can be said with a large degree of confidence that unconscious MAMs will not have freedoms or rights, and will be expected to just follow orders as a result of the purpose they are created for.

When examining IHL alignment, the chapter has shown that the current code of IHL ethics does not align with MAM decision making, both from the perspective of MAMs making decisions for humans, and IHL ethics acknowledging and accommodating IHL personhood, autonomy, and moral agency. IHL ethics presently only focuses on human decision making,

[679] J. Basl, 'Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines,' (2014) Journal of Philosophy and Technology, 27(1), 79-96. 2014 < https://philpapers.org/archive/BASMAM.pdf > accessed on 10th December 2018.

autonomy and welfare, hence does not recognise other conscious entities. Developing frameworks that address accountability, moral agency, and the protection of IHL ethics and values in the context of MAMs is essential to ensure that the ethical foundations of IHL remain intact. However, we must not be speciesist and overlook the protection. Ultimately, as we navigate the intersection of technology and warfare, a commitment to preserving human dignity and ethical conduct must remain at the forefront of our efforts to shape the future of armed conflict.

For MAMs, the chapter has shown that free will and being in the setting of war is complex and challenging. How much free will a soldier is able to employ when he has the responsibility of his comrades, is debateable and there is likely to be a degree of external (e.g. organisational) coercion in play. Presently, MAMs are in limited use and all MAMs are under human control, without any form of consciousness or free will, thus relying exclusively on the intrinsic control of their operators, and acting under the operators values. However, MAMs could be used to assist in exploring or recommending other courses of actions that are not initially apparent to the human commander, supporting the principles of proportionality and necessity. Further, MAMs could make combat safer for those who accidentally get caught up in the fighting (e.g., civilians). However, MAMs could lower the threshold for entry into war and if an error is made by a fully autonomous weapon, then the devastation and consequences are unbounded. MAMs will be able to process and understand conflict and the frontline more efficiently than humans ever have the ability to. Working side-by-side with humans will allow for extra checks and balances.

Building on this, it raises the important question of how we expand IHL to protect them, which is argued here as the true value alignment problem (TVAP).

If IHL values and principles will ever allow for life and death decisions to be made by an MAM, with no human control, is not a matter this chapter has decided on. However, the research highlighted that humanity is more likely to feel more wronged if a human is killed by any MAM, than by another human, particularly if there is an accountability gap. On the other hand, it has been highlighted that MAMs lack of emotions will ensure their judgement is not blurred and the frustration but could have a desire for self-preservation.

Ethical theories and principles will need to be adapted or created to address the novelty and challenges of AM/MAMs. No conscious entity should be ignored or denied basic rights, and it seems absurd to the author that IHL would intend this. MAMs introduce significant risks and challenges into the already complex environment of the battlefield. Thus, IHL will need to be developed and expanded to accommodate MAMs. Our values will need to be reviewed and expanded, resulting in amendments to IHL scope, which is viewed here as the TVAP. Very careful consideration should be given to deploying them, along with the justification to them of why they have been designed for such purpose. It will be interesting to see if MAMs agree with our justification for war or even the mere idea of it. It is still to be determined if they will have the right to decide to engage in conflict and even which State/side they choose to support.

6. RA4: To understand Whether MAMs Truly Align to the principles of IHL

6.1 Introduction

As highlighted in previous chapters, IHL has been designed to ensure human-to-human military action be undertaken with as little devastation and harm as possible. It places value on all human life and wellbeing, regardless of what side they are on. RA1 highlighted that when values are aligned between two fighting people, it is frequently found that they implicitly or explicitly come to an understanding of the limits of their warfare. Yet when enemy States clash due to differing values, religious beliefs, status, race, or language, it can feed a view that that each other is "less than human", and as such war conventions are seldom applied. Whilst MAMs may not be attached to values the same way that humans are, they are still an asset of a specific State, so will be indoctrinated with the values and cultures of that State. Consequently, this highlights the necessity of ensuring MAMs are developed and trained consistently throughout all States to adhere to International IHL and uphold the values to avoid liability issues.

IHL presently regulates insentient weapons and the force allowed, with all present day weapons being under the control of a human. However, MAMs, held here as having personhood and autonomy, will challenge this. Not only will there be an expectation that MAMs will demonstrate compliance with IHL towards humans and our values, but it is argued here that humans will need to review the principles under IHL with extending them to protect MAMs in mind, argued here as the true value alignment problem (TVAP).

Indeed, RA2 stresses that If we truly want MAMs to respect, uphold and see the value that we see, then we need to be an advocate for the benefits and proactively extend them to MAMs; the true value alignment problem (TVAP). To stress, it is argued here that the TVAP concerns how we recognise, accommodate, and protect the rights and views of MAMs under IHL. Indeed, RA1 argued that addressing these issues requires not only a rethinking of legal accountability and control mechanisms but also a deeper philosophical examination of the moral status of conscious entities and their rights in the context of warfare; the true value alignment problem (TVAP).

This chapter follows on from Chapter 5 and applies the IHL theories and discussion to MAMs.

6.2 IHL and Accountability for MAMs

For IHL, clear accountability is vital, as it ensures justice, deters violations, and facilitates compliance to the IHL aim of balancing legitimate military action with the "humanitarian objective of reducing human suffering, particularly among civilians"[680], via the core principles of distinction, necessity, proportionality, and humanity. Research aim 3 emphasised that responsibility and accountability are key foundational aspects of ethics, the VAP and thus attributes AM/MAMs should possess. MAMs possessing these attributes is asserted here as laying the foundations for being recognised as having autonomy and being moral agents,

[680] British Red Cross, 'International humanitarian law' (British Red Cross, 2024) < https://www.redcross.org.uk/about-us/what-we-do/protecting-people-in-armed-conflict/international-humanitarian-law#:~:text=Humanity%20forbids%20the%20infliction%20of,accomplishment%20of%20legitimate%20military%20purposes > accessed 4 September 2024.

which is stated here as ethical and further both address the VAP but manoeuvres the conversation and focus to the TVAP.

How MAMs will align to the accountability requirements for just their use, raises fundamental ethical and moral challenges beforehand, for example, if humankind's values and principles will ever allow for life and death decisions to be made by a MAM, with no human control. Cultivating their intrinsic control and imparting our values for military environments will be challenging, complex and risky, and we will have to wait to see if their decisions truly do align to our expectations of IHL protection, or if their interpretation and/or understanding differs. An expert speaker at the meeting on 'Autonomous weapon systems: Technical, military, legal and humanitarian aspects', held on March 2014 in Geneva, Switzerland, highlighted that "although moral sentiment and ethical judgement are not specified in the law and should not be confused with the law, these ethical elements are often used as a basis for formulating legal rules"[681], which leads to accountability and has already been discussed as challenging for MAMs. For example, moral judgement is considered to lie behind the assessment of whether a weapon could cause unnecessary suffering. Similarly, the Martens Clause exemplifies a moral structure where lethal force should not be used even against lawful targets, unless it is determined a necessity to kill. Speakers at the March Geneva meeting debated that IHL rules overseeing the behaviour and conduct of conflict have an influence over humans using human judgement, but will not have the same influence over MAMs, for example, due to the psychological aspects or the limitations of human biology, which is a view agreed with here. It has been argued in Research Aim 3, that absence of human biology and

[681] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

psychological aspects, could lead to better decision making by MAMs, resulting in better outcomes. This also highlights the true IHL value alignment problem (TVAP) of humans applying judgement and influence when considering MAMs, and what behaviour and conduct is lawful for and against them, considering their unique limitations. For example, interfering or disabling their sensors, or exposing them to corrupt data.

The research highlights that humanity is more likely to feel more wronged if a human is killed by an MAM, than by another human, regardless if the human would take the same action to kill. This can lead to a perception of injustice and the dehumanisation of war. Indeed, an expert at the March Geneva Meeting raised a question over the consequences of overriding the right to life via a piece of software and if it is morally right to do so. Removing a human from such a momentous decision, is seen by the author as being potentially inappropriate within non-military settings, e.g., a children's hospice. Yet, the author accepts that life ending decisions would be made by MAMs, as they would be deployed to take action that includes killing. The speaker also stressed that, "with increasing 'dehumanization of warfare' we may lose responsibility and moral accountability, as well as our ability to define human dignity."[682] This hints of a possible value alignment challenge within the human to human application of IHL, which is concerning in itself, and could be complicated further with the use of MAMs. The speaker was not supportive of handing over decisions to unconscious MAMs, and accentuated that it would be irresponsible to lose our oversight and remove ourselves from the loop, "since morality requires meaningful human supervision of decisions to take life."[683]

[682] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

[683] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

It is agreed with here that unconscious MAMs should not make significant decisions without human oversight. However, MAMs (being conscious) could make decisions that are more beneficial and cause less harm, due to the range and amount of data they can process within milliseconds, compared with a human solider. Again, they could make better decisions s discussed in Research Aim 3.

Accountability for severe IHL infringements performed by MAMs elicit issues around potential 'accountability gap' and consequently lack of extrinsic control, as discussed in chapter 3. The question over how a human can be held responsible for a MAM which they have no control of remains.[684] This is a major blocker for their use, and a gap in current thinking, as the risk they introduce and pose is not sufficiently understood and could be significant in terms of scale and impact. A theme falling out of the literature review, is the suggestion of a phased approach to the introduction of any new technology into military deployment, which would benefit MAMs, as it would build trust in their capabilities. Potential challenges around malfunction, errors, along with deliberate malicious programming of MAM to breach IHL, could mean that responsibility is assigned to a person in different stages, such as programming, manufacturing, and deployment of the MAM, although the liability hand-off points will need to be made very clear. It is the view here that there needs to be consistency in the values shared and the interpretation of them at all hand-off points, to ensure the MAM's training and development throughout reinforces IHL values rather than reinforce the VAP.

[684] Synectics, 'Evolution of Machine Learning' (Synetices, 2018) < http://www.smdi.com/evolution-machine-learning > accessed 3 March 2020.

Chapter 3 raised the question as to whether manufacturers will be willing to accept the liability for MAMs in environments where they have no control or influence, and that are often classified operations. This is seen here as playing a vital part in forming legislation, closing the accountability gap, and ensuring MAMs align to IHL, which establishes certainty, supports justice, and upholds the rule of law. In complying with the IHL principles, another question arises regarding if a MAM would be able to determine actions to take and make complex decisions to the same standard of a human? Indeed, speakers and participants at the Geneva conference proposed that MAMs, "should be held to a higher standard of performance than humans, partly because the public would be even less tolerant of war crimes committed by autonomous weapon systems than if they were committed by humans."[685] This is viewed to conflict with the possibility of MAMs having rights as they will be held to a higher standard than humans and will consistently have to perform at this high standard to maintain their status. It is questioned here again, following the discussion in Research Aim 3 (RA3), if this is equitable, and certainly underpins what this thesis argues here is the TVAP. To note, Research Aim 3 discussed not measuring MAMs by the standards of a 'reasonable person' from day one of their 'life', as it is held here that they will require a period of development and learning. As a result, RA3 argued that MAMs should not be deployed until they have had some degree of assessment to test their decision making and to ensure they act appropriately.

[685] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

6.3 Fundamentals of IHL and the True Value Alignment Problem (TVAP)

6.3.1 Recognising MAMs Personhood

Research Aim 2 showed that when looking at consciousness, philosophers Hegal[686] and Kant[687] link morality to ideas of consciousness, personhood, free will and rationality, which the author argues underpins the TVAP. The IHL assumes a person has personhood, so does not need to specifically mention or refer to personhood. Personhood is explored in Research Aim 3, but as a summary, personhood encompasses several attributes, including consciousness, communication and self-consciousness.[688] Of key importance, Research Aim 3 highlighted that Kitwood argues that personhood ought to be treated with profound respect. At the core of Kitwood's theory is the moral concern for 'others,' which is TVAP with regards to MAMs. MAMs will first need to have their personhood acknowledged and recognised. In gaining personhood status, the question arises as to how IHL will accommodate MAM personhood, which is argued here as a key factor of the TVAP. Applying the components of personhood to MAMs is multifaceted as firstly MAMs will be developed for the sole purpose of serving in the military, thus devoid of choice. One ponders if IHL will revise and extend the provisions under Article 51 to protect MAMs from forced conscription.[689] The revision of Article 51 could include expanding the pressure and propaganda protection to include forbidding a State form only allowing a MAM access to data that supports the State's view on military action and/or the enemy (e.g., propaganda). There could also be a revision of the prohibition of those under 18 years of age, along with protections for those over 18 who are

[686] G W F Hegel, 'Philosophy of Right' (2001) Transition (Vol. 1), Kitchener, Batoche Books Limited.

[687] I Kant, *Fundamental Principles of the Metaphysic of Morals*, (1785) 1949th ed. New York, NY: L. A. Press, Ed.

[688] D C Dennett, 'Conditions of personhood' [1976] The Identities of Persons, ed A. O. Rorty Berkeley, CA: University of California Press.

[689] Convention (IV) relative to the Protection of Civilian Persons in Time of War. Geneva, 12 August 1949, Article 51, which states ""The enlistment of other protected persons on the occupied territory is not prohibited, although no pressure, propaganda or any type of coercion may be employed to secure voluntary enlistment."

only deployed "on work which is necessary either for the needs of the army of occupation, or for the public utility services, or for the feeding, sheltering, clothing, transportation or health of the population of the occupied country."[690] As a result, it is mused here if MAMs who are newly switched on and still developing in the military environment, could be prohibited from the battlefield until such a time as they are assessed to understand the impact and consequences of deployment, and agree to be deployed. Whilst the author would welcome such protections for MAMs, there is concern that State's will not wish for this, as they may see MAMs as a substitute for human soldiers and so reducing causalities.

Research Aim 3 argues strongly that AMs and MAMs will eventually meet the requirements for personhood and subsequently be a moral agent, therefore recognising and assigning moral agency will be fair and equitable. As a result of becoming a moral agent, a MAM could be wronged under IHL, if the ethical considerations of IHL do not take into account the life and value of MAMs; the TVAP. Further, Research Aim 3 stresses the need for awareness of the influence and control of ANNs on MAMs, which is argued here as key facet in the TVAP and recognises their uniqueness, personhood, and autonomy.

Drawing from chapter 5, MAM's personhood status raises questions over the alignment with IHL, specifically:

a. Conscious MAMs – A question arises if MAMs should ever be deployed, especially as we are choosing the fate of a conscious entity. Are the risks too

[690] ICRC, 'Article 51 - Enlistment. Labour' (ICRC.org, 2024) < https://ihl-databases.icrc.org/en/ihl-treaties/gciv-1949/article-51> accessed 4 September 2024.

great for both the MAM and humanity? It is the view here that a conscious entity with personhood should not be forced to be deployed in battle without agreeing.

b. Self-Motivated – MAMs will interpret and experience the environment their own way via their learning, although they will have to follow orders and adhere to the chain of command. This may lead to them acting differently from expected. Self-motivation for MAMs is particularly complex as they will be designed and built with the sole purpose of service within the military. As previously highlighted, humans decide to join the military, with awareness of what it is likely to entail and experience of non-military life. This will not be an option for MAMs, therefore it is argued here that it is all the more important for their rights to be considered and upheld; A key part of the TVAP argument raised here.

c. Reasoning – Whilst MAMs will have good reasoning power and be able to understand large amounts of information, it is unlikely that reasoning will be a trait expressively looked for in am MAM, due to the need to follow orders and the concerns over accountability. Nevertheless, MAMs could be used to assist in exploring or recommending other courses of actions that are not initially apparent to the human commander, and their strategic thinking could create a competitive advantage.

Currently the above points have not been addressed within IHL, and thus MAMs personhood is not recognised and protected. It is the view here that these questions and points must be resolved and safeguards implemented before MAMs are deployed for a State. Therefore, it is recommended here that MAMs are aware of what the intended purpose of their existence is and allowed to 'opt out' of serving or taking certain decisions (e.g., to kill). Research Aim 3 strongly argues that IHL principles should be amended and extended to MAMs and thus recognise and protect their value and 'life', which is considered by the author as the TVAP. This requires a thorough understanding of the impact of the IHL principles viewed from the point of view of a MAM. For example, understanding what would constitute suffering for a MAM, and not just viewing it from a human aspect. Supporting the protection for their life and the option to 'opt out' of actions and decisions, it is the view here that MAMs could elect to take a back seat role or request a human signs-off on their decisions. Further, there could be periodic reviews with MAMs, to ensure their continued agreement. This could also feed into potential protection revisions under Article 51.

6.3.2 Understanding MAM Autonomy within the Military

Again, the IHL assumes a person has autonomy, so does not need to specifically mention it. Yet, MAMs will need to first have their autonomy acknowledged and recognised. As highlighted in Research Aim 1, autonomy is also the key area where the TVAP becomes glaring obvious; IHL assumes those fighting as soldiers do so with consent and through their own choice. MAMs will not consent but be designed to be deployed, thus their free will will be inhibited.

MAM autonomy was discussed and assessed earlier in chapter 5. However, the main points relevant to MAMs and how they align with IHL, are stressed here:

- Agency – MAMs will not have as much agency as they will be created for serving in the military, which is unlike any human, they will not choose their own life path.
- Free Will – For Owen, free will centres around decision-making and the power to veto actions or "free won't". Applying this in the military context to soldiers, it is hard to identify any true free will or free won't being exercised, which will be exacerbated with MAMs. Owen's idea of 'free won't' is very much at odds in the military context, yet even more so for MAMs, who are likely to be sacrificed ahead, or in replacement, of humans. Indeed, it is this freedom not to act (Owen's 'free won't') that is of most interest with MAMs along with their 'freedom to act'. As a result, it is argued here that this is the pinnacle of the IHL TVAP. In reality, the idea a MAM could refuse an order is difficult, as the military environment is the only environment they would ever have operated in or have significant experience of, thus they would have no experience or context of any other environment/setting. As a consequence, MAMs will not be able to frame their actions and decisions with insight and experience of other environments. Further, Hooker's[691] view that murder would be contrary to autonomy and suggests that we may have a duty to repair and/or rehabilitate MAMs rather than destroy them, is supported here and supports the IHL principle of humanity. The idea that we will simply be able to pre-programme much of MAM behaviour and even implant any cultures and/or personalities we wish, is considered here as interfering with the MAM's autonomy.

  How much free will a soldier is able to employ when he has the responsibility of his comrades, is debateable and there is likely to be a degree of external (e.g.

[691] J Hooker, 'Autonomous Machines Are the Best Kind, Because They Are Ethical', (2016), Carnegie Mellon University < http://public.tepper.cmu.edu/jnh/agencyPost2.pdf > Accessed on 14 December 2017.

> organisational) coercion in play. For example, an order given to enter an enemy location and resulting in a risk to the soldier's life; Does he really enter the location with free will? Does he feel pressure to follow the orders and feel pressure from the troop? Could a bomb disposal expert really decide that he doesn't want to risk his life defusing a bomb that could kill hundreds?

If an MAM does meet the definition for free will, then it is held here that we must recognise it has free will or we could end up in a speciesist position, risking discrimination and ignoring welfare for the belief of our own uniqueness. This is further viewed here as unethical and underpins the author's assertation of the TVAP. Should conscious MAMs be deployed into the battlefield, then it is further argued here that they should be afforded the same autonomy as intended with mission command and not be expected to behave similar to the WW1 troops, who were ordered over the trenches. What if a MAM does not want to take an action because the risk of harm to itself is too great? Would they be at greater risk of external coercion and punishment? We currently see technology as subservient and disposable, so protecting a soldier's life over a MAM's could naturally follow without hesitation, if provision is not made within IHL. This is again a core facet of the TVAP and it is the view here that this would not align with to the ethos of IHL. Thus, it should not automatically apply that a human soldier's life holds greater value and moral agency, as discussed in chapter 5, and in fact a MAM could be protected by human soldiers.

Consequently, because of the limits on MAM autonomy and free will, it is argued here that the TVAP becomes amplified. It is the argued here that by recognising and acknowledging MAMs autonomy, along with extending IHL to protect them, it actually helps us to take a fresh look and a reviewed understanding of what it is to be human, our values, and ultimately makes us more humane. This surely is in the best interest for staying true to the ethos of IHL and aligns to the sacrosanct principles. Indeed, it shows how we grow and evolve as a species and society, and how we empower autonomy over decisions and choices, e.g., gay marriage, transgender, MAMs 'life'.

6.3.3 TVAP: Application of the Principle of Distinction to MAMs

It is the view here that to align to the principle of distinction, a MAM and a soldier will both need to be trained to make the distinction between a MAM and a non-MAM alongside civilians and soldiers. Presently, MAMs are expected to distinguish between civilians and soldiers, but no consideration have been given within IHL to distinguishing between MAMs and unconscious MAMs, which again feeds into the TVAP argued here. In addition, just as it would be viewed unethical to kill a soldier who has surrendered, it should also be extended to include MAMs that have surrendered. Thus, military leaders will need to be mindful that a MAM has the right to surrender in accordance to the LOAC[692], just as a human soldier has. Consequently, the MAM would be considered hors de combat, which aligns with IHL, and as a result, feigning surrender would be viewed as an act of perfidy and so forbidden.[693]

Deploying MAMs will resent risks due to their lack of perception and ability to understand human dynamics and subtleties. For example, how could they tell if someone was acting under duress? Could it be considered inhuman for a detainee to be looked after by a MAM

[692] Law of Armed Conflict

[693] ICRC, 'Surrender' (ICRC.org, 2024) < https://casebook.icrc.org/a_to_z/glossary/surrender#:~:text=In%20international%20law%2C%20an%20isolated,perfidy%20and%20is%20therefore%20forbidden > accessed 4 September 2024.

and not a military medic or soldier? On the other side, if MAMs have personhood, they should be protected and treated with respect as human soldiers are. Indeed, we protect service animals[694], so it would not be too far a leap to expect protection for MAMs.

6.3.4 TVAP: Application of the Principle of Proportionality to MAMs

As highlighted previously in chapter 3, the force used should be 'proportional' or 'appropriate' and no more than is required to ‘win’, which can lead to a wide interpretation. It certainly creates a grey area where a human soldier’s intrinsic control could be stretched, compromised or misaligned.

Keeping a human in the loop and in control with such advanced autonomous machines, may be considered an oxymoron. The pace of MAM development should be seen as a real threat as technology races are already established with our current use of technology, for example, LLMs[695]. Thus, we could face an MAM arms race as the technology advances and hostile States develop and even deploy MAMs into the battlefield, which will change the dynamic of the battlefield. Further, the deployment of MAMs may be disproportionate to a State that does not have such capability, which would not align with the principle of proportionality.

As a core component of the TVAP, as previously discussed in chapter 5, is that harm and suffering will be different for MAMs, so military decision making and operations will need to consider this when planning. However, the application of the proportionality principle is not always clear-cut and sometimes a method of attack that would reduce the risk to civilians

[694] Animal Welfare (Service Animals) Act 2019.

[695] Large Language Model

may increase the risk to attacking forces, including MAMs. Unfortunately, other States are not as ethical and play on our ethical status and challenge us, for example, during the Gulf War of 1991, where Iraq deliberately placed military objectives near and around protected objects (e.g., mosques, medical facilities, and culturally significant property).[696] More recently, during the Israeli – Hamas conflict, Israel deliberately militarised civilian objects, and turned schools into military bases during .[697] Thus, we would need to protect MAMs against becoming a target.

In addition, the risk of how a MAM would evaluate such a situation and calculate the risk versus benefit, is at present unclear due to the technology still developing towards machine consciousness. MAMs, both conscious and unconscious, will be equipped with sensors that can observe and interpret the frontline environment better and quicker than any human, which includes "ongoing technological advances in electro-optics, synthetic aperture or wall-penetrating radars, acoustics, and seismic sensing, to name but a few."[698] It is reasonable to expect that, based on the research, MAMs will be able to process and understand conflict and the frontline more efficiently, but a question arises as to whether we would actually want MAMs to make such critical decisions?

However, the benefits of humans working alongside MAMs will allow for extra checks and balances; "When working in a team of combined human soldiers and autonomous systems as an organic asset, they have the potential capability of independently and objectively monitoring ethical behaviour in the battlefield by all parties, providing evidence and reporting

[696] US Department of Defense, Conduct of the Persian Gulf War, Final Report to Congress (1992) (Department of Defense Report) 613.

[697] Euro-Med Human Rights Monitor, 'Gaza: Israel deliberately militarizes civilian objects, turns schools into military bases' (Euro-Med Human Rights Monitor, 2024) < https://euromedmonitor.org/en/article/6296/Gaza:-Israel-deliberately-militarizes-civilian-objects,-turns-schools-into-military-bases > accessed 4 September 2024.

[698] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

infractions that might be observed. This presence alone might possibly lead to a reduction in human ethical infractions"[699], which aligns to IHL. Nevertheless, whist cognisant of human values and aiming to reduce human ethical infractions, there is no consideration for MAM ethical infractions, which the author asserts underpins the TVAP. Further, we must not view or default to MAMs being our minders and expecting them to call out or correct our bad behaviour, especially if we do not recognise or correct bad behaviour towards them. We will still be accountable for ourselves and should not expect MAMs to parent us.

As highlighted at the March Geneva Meeting[700], there are counter-arguments, such as determining responsibility for war crimes involving MAMs, along with the possible, "lowering of the threshold for entry into war, the military's possible reluctance of giving robots the right to refuse an order, proliferation, effects on squad cohesion, the winning of hearts and minds, cyber security, proliferation, and mission creep."[701] Nevertheless, MAMs could overtake humans in the frontline when assessing obedience to IHL and yet not compromise mission performance, and without the need to replicate all human moral capacity. Understandably apprehensions exist and need to be addressed, for example, target discrimination and the detection and distinction of those hors de combat.[702] Indeed, the pace of the frontline is overtaking the capability of soldiers to make good balanced decisions under the stress of combat. MAMs could make combat safer for those who accidentally get caught up in the

[699] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

[700] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

[701] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

[702] Meaning 'out of combat.' It is a French term used in diplomacy and International Humanitarian Law (IHL).

fighting (e.g., civilians), but in doing so, they wll be placed in situations where they could be harmed or destroyed. Thus, respecting their 'life' and their potential sacrifice through extending IHL values, is argued here as critical to addressing the TVAP as identified and argued by the author.

When looking at proportionality in relation to protection of MAMs, it must be considered as to what force is proportional and appropriate to use against MAMs, which is part of the TVAP. For example, would denying power or access to data be proportional and appropriate? These will need to be determined before deploying MAMs, so all States are in agreement and incorporated into their military policies.

6.3.5 TVAP: Application of the Principle of Military Necessity to MAMs

All weapons in use by the British military are lawful as determined by the BA, however they must be used in the correct way and no alterations or changes should be made to them.[703] It would probably be considered unethical to use a machine gun against an enemy who is only armed with bats, but this could be the only weapon available. Thus, the first question to arise is, are MAMs necessary to meet a legitimate aim? Subsequent questions arise as to will MAMs be able to understand and justify military action? Could their default, gained through their learning and from only being exposed to the military environment, be to go to war?

[703] Ministry of Defence, 'A Soldier's Guide to The Law of Armed Conflict' (2005) Ministry of Defence <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/619906/2017-04714.pdf > accessed 2 October 2018.

Regardless of how we choose to engage in conflict, MAMs must obey the rules and honour Just War Theory (JWT), just as any other military personnel. Yet, necessity needs reviewing in light of MAMs and the appropriate protection and safeguards in place for them; The TVAP.

The MoD are clear that any, "killing or wounding the enemy by treachery is forbidden,"[704] and that:

> "a white flag of truce may be used to signal a wish to talk to the enemy. The side using the white flag must stop fighting and indicate a wish to communicate. Both sides must then stop fighting. Abuse of the white flag is treachery."[705]

Whether a MAM could fully understand and appreciate these complex, emotive and subtle rules, which are entrenched in history, and closely linked to our human past, is still to be determined. Further, when looking at the TVAP, there is a need to look at adapting such rules to ensure the MAM is not harmed and does not become a target itself, and perhaps even wave a white flag?!

6.3.6 TVAP: Application of the Principle of Humanity to MAMs

IHL has limiting suffering and harm are at the core, whilst protecting those involved in conflict, both directly and indirectly. Humanity underpins all the other principles, and encompasses

[704] Director General, 'JSP 301: Aide Memoire on the Law of Armed Conflict' (Government Publishing, 2010) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/902747/dcdc_legal_aide_memoire_law_armed_conflict_jsp381.pdf > accessed 3 March 2017.

[705] Director General, 'JSP 301: Aide Memoire on the Law of Armed Conflict' (Government Publishing, 2010) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/902747/dcdc_legal_aide_memoire_law_armed_conflict_jsp381.pdf > accessed 3 March 2017.

several notions, such as "suffering is universal and requires a response: it cannot be met with indifference"[706], "respect for human dignity is paramount"[707], and protection for human life and health.[708] Research Aim 3 highlights that suffering could take the form of reduced processing capacity due to being denied a software upgrade, for example. This type of suffering (e.g., reduced processing power) could be 'real' for the MAM, even if not considered as suffering to us. Understanding and acknowledging this is argued here as a vital aspect of the TVAP. As a result, extending IHL values and principles to MAMs is argued here as the TVAP and will ultimately test us to understand how much value and weight we place on our values; The 'Golden Rule'[709]. Extending our vales could increase humanity and make us more human.

In upholding the principle of humanity, Arkin asserts that unconscious MAMs, if managed properly, could result in greater adherence to IHL by MAMs than from using human soldiers. [710] Further, Arkin caution's that:

> "It is unacceptable to be 'one war behind' in the formulation of law and policy regarding this revolution in military affairs that is already well underway. The status quo with respect to human battlefield atrocities is unacceptable and emerging technology in its manifold forms must be used to ameliorate the plight of the non-combatant."[711]

[706] ICRC, 'The Fundamental Principles of the International Red Cross and Red Crescent Movement' (2015) 4046/002 08.2015 5000 < https://www.icrc.org/sites/default/files/topic/file_plus_list/4046-the_fundamental_principles_of_the_international_red_cross_and_red_crescent_movement.pdf > accessed 4 September 2024.

[707] ICRC, 'The Fundamental Principles of the International Red Cross and Red Crescent Movement' (2015) 4046/002 08.2015 5000 < https://www.icrc.org/sites/default/files/topic/file_plus_list/4046-the_fundamental_principles_of_the_international_red_cross_and_red_crescent_movement.pdf > accessed 4 September 2024.

[708] ICRC, 'The Fundamental Principles of the International Red Cross and Red Crescent Movement' (2015) 4046/002 08.2015 5000 < https://www.icrc.org/sites/default/files/topic/file_plus_list/4046-the_fundamental_principles_of_the_international_red_cross_and_red_crescent_movement.pdf > accessed 4 September 2024.

[709] "The Golden Rule", Matthew 7:12, New Testament.

[710] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

[711] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017

Arkin's views are of great interest here, as 'one war behind' could be viewed from a MAM consciousness perspective. This could be viewed as not planning for the the violations and atrocities that could result, or having any regard for, MAM consciousness, along with all the rights and duties that bestows, which is argued here as a fundamental part of the TVAP.

Martens Clause is an interesting component of humanity and whilst it is highly unlikely that Fyodor Fyodorovich Martens had MAMs in mind when introducing the clause for the first time during the opening of the 1899 Hague Convention, it is argued here as extremely relevant for MAMs and ensuring they will "never find themselves completely deprived of protection."[712]. His ideology has stood the test of time and been long upheld, hence it is the view here that the protection for "persons" could be extended to include personhood and thus recognise MAMs, which helps address the TVAP.

Military culture and norms are exclusive to the military environment, which sets to dehumanise who the State deems as the enemy and allows for killing. The military have already identified potential benefits of using MAMs, including:

> "a reduction in friendly casualties; force multiplication; expanding the battlespace; extending the soldier's reach; the ability to respond faster given the pressure of an ever-increasing battlefield tempo; and greater precision due to persistent stare (constant video surveillance that enables more time for decision-making and more eyes on target)."[713]

[712] ICRC, 'Martens Clause' (ICRC.org, 2024) < https://casebook.icrc.org/a_to_z/glossary/martens-clause > accessed 4 September 2024.

[713] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

This makes a case for the development and deployment of MAMs from a not only a military efficacy and economic viewpoint, but also for the application of IHL. Despite this, how metrics would be created to validate such benefits is unknown, for example, it is unlikely in the complex throes of war, that it would be easy to tally the lives lost because of a human decision verses the lives lost through MAMs decisions. It should be reiterated that unfortunately past and present human behaviour in conflict, concerning the adherence to legal and ethical obligations, is dubious at best. Mankind does not have a great record of ethical behaviour during conflict, as the case of 'Marine A'[714] proves[715], and again, MAMs could potentially lead to a reduction in non-combatant deaths and casualties. Indeed, their adherence to the values and principles of IHL could surpass our own, and underlines why the author advocates for addressing the TVAP; They need to see the value and benefit for themselves.

Humans "feel the emotional weight and psychological burden of choosing to take away the life of other human beings…Empathy can act as a check on killing, but only if humans have control over whom to target and when to fire,"[716] they ignore the fact that war, but its very nature, is intended to kill someone, albeit viewed as 'the enemy' thus justified and accepted. Is it fair to put the burden of taking human life on another human, which often results in Post-Traumatic Stress Disorder (PTSD) for military personnel? A question for future consideration is if military personnel would really feel less burdened by handing over responsibility to MAMs or would they end up feeling guilty for the MAMs actions? If an MAM is proved to have

[714] Marine A is Alexander Blackman, the Royal Marine sergeant jailed for killing a wounded unarmed Taliban fighter. The case is discussed in the legal chapter.
[715] discussed in the Chapter 4, Law.
[716] Human Rights Watch, 'Killer Robots and the Concept of Meaningful Human Control' (2016) <https://www.hrw.org/news/2016/04/11/killer-robots-and-concept-meaningful-human-control > accessed on 6 November 2016.

psychological awareness, would it be fair to burden them with the decision? This is why it is argued here that we need to adapt our values to consider them, their best interests, and welfare; The TVAP.

Whilst MAMs will avoid the human psychological problem of 'scenario fulfilment,'[717] which can lead, "to distortion or neglect of contradictory information in stressful situations, where humans use new incoming information in ways that only fit their pre-existing belief patterns"[718], and is believed to be responsible for the shooting down of an Iranian passenger airliner by the USS Vincennes in 1988, MAMs will act on the information/data they have and what they have learnt from their leaders or troop. A State's military will ultimately be their 'teachers', thus MAMs will need to have teachers that can show them the correct behaviour to ensure MAMs do not breach laws, policy and ethical boundaries and have safeguards to avoid exploitation by a leader/troop for their own agenda.

6.4 Summary

IHL values and principles, borne out of the atrocities of yesteryear, can be used to remind ourselves to be humane and value life, whatever form that takes. This chapter has explored how MAMs align to IHL and has shown, that whilst they could align to our IHL values and subsequent protections via development and training, no consideration has been given for

[717] Psychologists hold this occurs when a person is under pressure. In this context, a military person would undertake a training scenario, in the belief it is reality, whilst ignoring information that opposes the scenario.

[718] ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017.

the protection of MAMs. Therefore, IHL is not aligned due to mankind's needs far outweighing MAM's needs, recognition or their impending life, personhood, and autonomy, which is argued here as the TVAP.

The deployment of MAMs into conflict represents a value alignment problem (VAP) under IHL because it upends the assumptions of agency, consent, responsibility, and adherence to humanitarian principles that are fundamental to current legal and ethical frameworks. Addressing these issues requires not only a rethinking of legal accountability and control mechanisms but also a deeper philosophical examination of the moral status of conscious entities and their rights in the context of warfare; the true value alignment problem (TVAP). Nevertheless, the author stated that Martens Clause is extremely relevant for MAMs and could be used to ensure they "never find themselves completely deprived of protection."[719].

The chapter again argued for MAMs to have their autonomy recognised, and has highlighted, that MAMs, unlike human soldiers, will be created for the purpose of being a MAM and will therefore have the fundamental element of free will curtailed. Thus, it is strongly held here that, because of their limited choice and freedom, MAMs should have their rights recognised and protected at the earliest.

Deploying a conscious entity into a complex and human created environment such as war, and ordering it to fight, sits uncomfortably with the author. The life of a MAM will be governed by the military of the State. Regardless, we should be looking at mitigating risks for all MAMs now, as hostile States are developing their MAM technology at pace and it would be

[719] ICRC, 'Martens Clause' (ICRC.org, 2024) < https://casebook.icrc.org/a_to_z/glossary/martens-clause > accessed 4 September 2024.

extremely dangerous to be disadvantaged and ‘one war behind’. One war behind here is seen here as including ignoring the personhood and autonomy of MAMs, which includes expanding Article 51 so MAMs are not forced to be deployed.

Deploying MAMs will present risks due to their lack of perception and ability to understand human dynamics and understand subtleties, so teaching will be key, although a significant overhead. Both a conscious MAM and a soldier will need to be trained to make the distinction between a MAM and a non-MAM alongside civilians and soldiers. Further, military leaders will need to be mindful that a conscious MAM has the right to surrender, which will further complicate the battlefield.

It is the view here that techniques and methods will need to be revised with the deployment of conscious MAMs, as techniques and methods under proportionality will differ for MAMs. We will need to look at adapting these rules to ensure the MAM is not harmed and does not become a target itself. MAMs will also need to be protected from any unwarranted interference or removal of data and/or hardware without their consent as this would be unlawful. Further, MAMs will come under more scrutiny should they make an error, as it is likely humans will be less accepting of technology making mistakes than a human. MAMs could become a target themselves and therefore consideration needs to be given to the protections they are afforded.

This extension of IHL values to accommodate MAMs, is argued here as the TVAP and will ultimately test us to understand how much value and weight we place on our values; The

‘Golden Rule’[720]. The research has brought to the fore that the principle of humanity, which underpins IHL, is crucial. Extending this to MAMs could increase our overall humanity, yet it also raises difficult questions about the moral status and treatment of non-human entities. Indeed, how can we ask another conscious entity to align, respect and uphold to our values if we do not even recognise and value their existence and ‘life’?! The author asserts that ethical theories and principles will need to be adapted or created to address the novelty and challenges of MAMs. Our values will need to be reviewed and expanded, resulting in amendments to IHL scope beforehand. Very careful consideration should be given to deploying them, along with the justification to them of why they have been designed for such purpose. It will be interesting to see if MAMs agree with our justification for war or even the mere idea of it. It is still to be determined if they will have the right to decide to engage in conflict and even which State/side they choose to support.

It was shown that applying IHL principles like distinction, proportionality, and military necessity to MAMs is problematic. MAMs may be able to make more informed decisions, but removing humans from the loop raises concerns about dehumanising warfare. Accountability for MAMs is a major challenge, as it is unclear how to hold humans responsible for the actions of autonomous systems, which they do not fully control. This "accountability gap" is a significant barrier to the deployment of MAMs and could ultimately limit their use.

MAMs will come under more scrutiny should they make an error, as it is likely humans will be less accepting of technology making mistakes than a human. For now, we can take comfort

[720] “The Golden Rule”, Matthew 7:12, New Testament.

that a human, who we expect has our best interests at heart or at least has an objective process to follow (e.g., IHL), makes the final decision during conflict.

7. Discussion Chapter

7.1 Introduction

This chapter provides a summary and discussion of all the chapters within this thesis, drawing the research together and interpreting it in order to address the following research aims and make recommendations:

1. To explore and understand the legal landscape of AI and IHL.
2. Exploration to understand if machine consciousness has been designed to work with IHL.
3. Examination into whether the current code of IHL ethics aligns with machine consciousness decision-making.
4. To understand whether MAMs truly align to the principles of IHL.

7.2 Research Aim Discussion

7.2.1 Research Aim 1: To explore and understand the legal landscape of AI and IHL

The author drew out the quote from the literature, which states, "ethics may be simply described as 'the intrinsic control of good behaviour'. This contrasts with 'law' that acts as the 'extrinsic control of good behaviour'". The author views this quote as which as epitomising the VAP challenges and a profound statement of the manner and reach of control, and feeds into what the author has been termed the true value alignment problem (TVAP). The literature demonstrates that the development of autonomous machines (AMs) and military autonomous machines (MAMs) are disrupting our stable legal and ethical landscape. It is stressed throughout the thesis, that AM technology used in a non-military setting, has been researched and relevant, due to the technology development pathway being the foundation of MAMs.

Whilst AM consciousness is yet to be established, it is shown that the current English civil and criminal law does not lay a path for its creation and recognition, as it assumes a human is always in control and thus responsible. In addition, current English law does not make provisions for any safeguards regarding machine consciousness. It is highlighted that we should be mindful of the rights and duties we owe AM/MAMs, particularly under IHL, thus develop, or create, new legislation around this. It is further considered that ignoring a conscious entity is absurd, especially if we are responsible for its creation, and this could be viewed as speciesism. Therefore, it is highlighted that laws will need to be created or adjusted to embrace the development path of AMs, to acknowledge their advancement, and to address the legal challenges and risks they will introduce. A first step to doing this would be to recognise them as a legal person, which legislation can flow from.

It has been shown that the current product laws[721], researched to set the context of liability, are insufficient for conscious AM/MAMs. Following on, the focus moved onto criminal liability, as criminal liability has the highest evidential bar and the core elements (actus reus and mens rea) have only been established in humans. It has been found that AM/MAMs could meet the elements of criminal liability in the future, thus liability should shift. Further, there

[721] The Consumer Protection Act 1987 (CPA 1987) is central to product liability, with Parts II and IV covering criminal liability. Tortious liability common law claims are founded upon Donoghue v Stevenson [1932] AC 562.

are situations where military personal, and thus MAMs, could be tried for criminal liability,, in either the military or civilian justice system, due to acting outside of International Humanitarian Law, for example, the unlawful killing of civilians or prisoners of war.

Once AMs are conscious[722], they will be capable of setting their own goals and acting with free will, it would therefore be unjust to hold the manufacturer, owner, or operator liable for an entity that they had no control over. As a result, it is asserted that it will be appropriate for AMs to be criminally responsible in their own right and for MAMs to be held accountable for breaches of IHL. Consequently, a question arises if an AM/MAM will be afforded 10 years of development and 'maturing' time before being capable of being criminally liable. It is viewed by the author that this will be considered unacceptable and further, it is the asserted here that liability will be assigned to them from the moment they are switched on. As a result, it is held here that a gradual process of liability transference will provide comfort, build trust, and test the process. This approach allows for reviews, adaptions, and even to stop the transference, if required.

To address their development and transference of criminal liability, Hallevy's[723] three models are explored and shown to be a very attractive and a flexible solution, which could be tailored to human-AM criminal liability as well as AM-AM liability. Hallevy's models are, at the time of writing, the only future thinking criminal liability framework that aimed to accommodate the development pathway to AM consciousness. It is the view here that Hallevy's models would

[722] Consciousness will likely emerge from the synergy between communication, internal knowledge, external knowledge, goal-driven behaviour, and creativity as per the research of . Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

[723] Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

only need to be slightly adapted for English law, with the inclusion of recklessness specifically for the first 2 of his models. His third model guards against future conscious AMs and allows for AMs to be assigned criminal responsibility, along with allowing them to have rights and duties.

It is shown that a human soldier can commit a crime if they step outside the permitted action and/or in excess of it, and highlighted that this does not cover MAMs, creating a liability gap, again illuminating that there are significant gaps in the current legislation. Indeed, current legislation[724] does not regulate the development beyond that of an inanimate object or tool, and views MAMs as being entirely under the control of a human during its operation and deployment. Further, there are key questions over how an MAM will be created and trained to understand, interpret and thus, adhere to complexities and subtleties of IHL. As a result, it is the author's opinion that it could be considered unlawful and irresponsible to deploy MAMs into battle if accountability is not clarified, and the risks they pose not exposed and mitigated against.

The principle that 'accountability follows control'[725] is presently viewed with a human in mind, but could be applicable with MAMs, which could be held accountable for any breaches of IHL. However, this is unlikely to sit comfortably for many, as it could appear we are absolving ourselves of liability and any harm caused. As a result, it has been highlighted that for the British Army, they would need to consider their own responsibility in regard to decisions made

[724] Both English civil and criminal legislation.

[725] Ministry of Defence, JSP 815. Element 5: Supervision, Contracting and Control Activities (JSP 815) Ministry of Defence < https://assets.publishing.service.gov.uk/media/66e18438dd4e6b59f0cb2500/JSP_815__Element_5_Supervision__contracting_and_control_activities_v1.2.pdf > accessed 4 September 2024.

and their control of MAMs. Nevertheless, for there to be effective extrinsic control[726], it is shown that unconscious MAMs will stay under human control for many years due to the risks and political fallout of deploying a conscious entity without a human in the loop or 'meaningful human control'. Further, the ethical and legal arguments around creating a conscious entity with the sole purpose of serving on the battlefield may ultimately prove too complex and therefore only MAMs without consciousness or legal personhood may be deployed. This could result in States being required to decide if conscious MAMs can ever be deployed or draft new polices to include this beforehand.

The research emphasised the difficulty in programming MAMs to adhere to IHL principles, and every situation they may encounter. The principles of distinction, proportionality, and humanity[727] are complex for automated systems to interpret and act upon in dynamic conflict scenarios. However, MAMs will have advanced technology to aid human soldiers and could protect their lives and reduce the devastation. Thus, ultimately it will come down to a risk (both in terms of global power and legal), versus benefit analysis and the amount of risk and reputational damage the State is comfortable taking.

Through exploring the legal landscape and IHL, it is shown that there is a significant liability gap arising with AMs and MAMs. AMs/MAMs will challenge our laws and possibly push them to breaking point, therefore we must be proactive and avoid a liability gap wherever possible. Thus, if we want to encourage innovation and technology development, then we be clear

[726] A.B.A Majeed, 'Roboethics - Making Sense of Ethical Conundrums' (2017) Procedia Computer Science, Volume 105, 2017 < https://doi.org/10.1016/j.procs.2017.01.227 > accessed 19 March 2019.

[727] ICRC, 'The Fundamental Principles of the International Red Cross and Red Crescent Movement' (2015) 4046/002 08.2015 5000 < https://www.icrc.org/sites/default/files/topic/file_plus_list/4046-the_fundamental_principles_of_the_international_red_cross_and_red_crescent_movement.pdf > accessed 4 September 2024.

where liability lies. Further, the deployment of conscious autonomous military machines into conflict represents a value alignment problem (VAP) under IHL because it upends the assumptions of agency, consent, responsibility, and adherence to humanitarian principles that are fundamental to current legal and ethical frameworks. Addressing these issues requires not only a rethinking of legal accountability and control mechanisms but also a deeper philosophical examination of the moral status of conscious entities and their rights in the context of warfare; the true value alignment problem (TVAP). Indeed, of most concern is that MAMs will not consent but be designed to be deployed, being the pinnacle of the IHL true value alignment challenge.

7.2.1.1 Research Aim 1 Recommendations

- Adapt and Implement Hallevy's "Perpetration-by-Another Virtual Responsibility Model" and "Natural-Probable-Consequence Virtual Responsibility Model" immediately into English law. They provide an effective framework to address the potential problem of people committing crimes via AMs and expecting to avoid liability.
- Prepare to implement Hallevy's "Direct Virtual Responsibility Model" as AMs develop greater consciousness and autonomy. This model holds the AM criminally liable if it can satisfy the requirements of mens rea and actus reus, which will be crucial as AMs advanced.
- Ensure the law keeps pace with AM development, rather than relying on outdated assumptions. The law should be ‘future-proofed’ to cover the challenges posed by unconscious and conscious AMs, rather than assuming AMs will always be under human control.

- Establish legal personhood status for AMs, defining their rights, duties and liabilities. This is necessary to properly accommodate and integrate AMs into the legal framework, rather than treating them as mere objects or tools.
- The author recommends legal systems begin preparing for MAM personhood and liability assignment, particularly for conscious MAMs that can make independent decisions, which will be crucial for future-proofing law and protecting human rights.
- Regarding military applications, ensure IHL is adapted to address the unique challenges posed by MAMs. This includes:
    - Developing robust testing, handover of responsibility, and legal review processes for MAMs throughout their lifecycle, not just initial development.
    - Exploring how MAMs can be programmed to uphold IHL principles like distinction, proportionality and humanity.
    - Establishing how to protect the "rights" of MAMs, with their own goals and decision-making, under IHL.

- Recognise the TVAP with regards to MAMs, as they will not have the same inherent consent and choice as human soldiers yet will deployed without choice. This is seen here as a fundamental tension in adapting IHL.
- Maintain human control and oversight of MAMs, while also allowing them appropriate autonomy. Striking this balance will be crucial to upholding legal principles and avoiding liability gaps.
- Ensure manufacturers, programmers, users and the military all have clearly defined responsibilities and liabilities regarding AMs/MAMs, to prevent exploitation of liability gaps.

These recommendations are key to proactively addressing the challenges AM/MAMs present, and a comprehensive, forward-looking legal framework is essential to ensure the responsible development and use of AMs and MAMs

7.2.2 Research Aim 2: Exploration to understand if machine consciousness has been designed to work with IHL

It has been highlighted that humans have always driven the need for more sophisticated tools to enhance our lives. The tools we have used over hundreds of years have been inanimate, under our control and beholden to us. AM development is another step in our tool development; however, AMs have the ability to surpass our knowledge and capabilities and introduce new risks, as a result of their 'black box' characteristics.

It has been shown that present day computers are subservient, without consciousness, under our command, and with our rights clearly outweighing theirs. However, it is argued that this may not always be the case, and we need to be prepared to accept this or simply stop the development of AMs. We need to define the humane treatment of AMs and consider questions such as at what point might we consider deletion of AMs algorithms as form of mass murder? Questions such as these raised in this thesis, highlight that the complexity, rapid pace and trajectory of AM development and will test our 'extrinsic' controls and standards. It has been highlighted that machine learning (ML)[728] and Deep Learning (DL)[729] advancements are crucial in developing machine consciousness. As a consequence, it is

[728] Synectics, 'Evolution of Machine Learning' (Synetices, 2018) < http://www.smdi.com/evolution-machine-learning > accessed 3 March 2020.

[729] Synectics, 'Evolution of Machine Learning' (Synetices, 2018) < http://www.smdi.com/evolution-machine-learning > accessed 3 March 2020.

asserted that this raises fundamental questions about free will, agency, and autonomy that have traditionally been discussed only in the context of humans and animals. Further, it is stressed by the author that there needs to be a universal scale for autonomy

It is the establishment of machine consciousness that entails understanding and aligning MAMs with human values and IHL principles. As a result, it is highlighted that machine consciousnesses and the value alignment problem (VAP) is considered as a result of the technological developments in AM/MAMs. In this context, the VAP relates to how we develop AMs to behave and act in accordance with human norms and values, which can be nuanced.[730] Further, it is shown that the 'problem' is amplified and intensified when looking at the values underpinning IHL, as the potential consequences and loss could be devastating.

Whilst designing and developing machine consciousness will be a major feat; it is actually recognising the implications and consequences for our treatment of AMs and MAMs that is asserted here as the biggest challenge and where the true value alignment problem (TVAP) arises. Indeed, establishing consciousness initiates the argument for recognising personhood. This subsequently leads to realising agency and free will, which constitutes autonomy. As a result, if we truly want them to respect, uphold and see the value that we see, then we need to be an advocate for the benefits and proactively extend them to MAMs, which it is argued here is the true value alignment problem (TVAP).

[730] Brian Christian, *The Alignment Problem: How Can Artificial Intelligence Learn Human Values?* (September 2021, Atlantic Books)

It is stressed by the author that machine consciousness has not been specifically designed with IHL in mind, and indeed the 'black box' characteristics with AMs support this and likely to cause challenges in determining and understanding the degree of value alignment. MAMs must interpret complex IHL principles like proportionality and distinction in unpredictable scenarios. However, achieving this in MAMs is still limited by present day AI, which lacks human situational awareness, empathy, and moral reasoning, it should be noted that nor is any human born to comply with IHL. Soldiers are educated and trained in IHL to understand and comply with the IHL framework hence, it is argued here that both the MAM and human brain start off as a blank canvas, and the argument of nurture versus natured argued for both. Yet, humans typically choose to enlist in the armed forces and thus come to identify and uphold IHL, of which MAMs will not be afforded such choice. Therefore, it is asserted that there will need to be an extensive period of training to defined and establish a MAM universal IHL standard, to understand the nuances and perhaps even a negotiation with MAMs as to why they should adhere to IHL. MAMs will have to be trained to recognise and understand the IHL compliant methods and means of warfare, in order to limit to suffering and injury, which will arguably be the hardest for a MAM to understand, as human suffering is unique due to psychological and biological factors, and hard to translate to MAMs, yet the author asserts that suffering will need to be established in relation to MAMs to ensure they do not suffer due to our ignorance. Thus, for MAMs to respect, align and uphold our IHL values, it is argued that we must demonstrate good conscience and equity and extend IHL and the inherent values, to them; This is argued by the author as the TVAP.

Despite the technology not being specifically designed with IHL compliance in mind, it is regarded as a requirement on the State to only deploy technology into the military environment that is deemed compliant; The State accepts the risk.

7.2.2.1 Research Aim 2 Recommendations

- Develop a standardised and universally accepted scale for autonomy.
- Draft design and safety standards for the development of conscious AMs.
- Undertake further research non-human consciousness and develop an ethical framework for monitoring and measuring AM consciousness, that is not solely based on human metrics.
- Develop a Comprehensive Framework for Machine Consciousness decision making and IHL Alignment:
    - Establish a collaborative, multidisciplinary community involving legal experts, ethicists, military strategists, and AI researchers to develop a comprehensive framework for addressing the challenges posed by machine consciousness decision making in the context of IHL, including addressing the TVAP.
    - This framework should provide clear guidelines, principles, and processes for designing, developing, and deploying MAMs, which can reliably uphold the core tenets of IHL (distinction, proportionality, necessity, and humanity), both in terms of MAM-to-human and human-to-MAM.
- Invest in Extensive Testing and Validation for MAM Decision-Making:
    - Implement rigorous testing and validation procedures to ensure MAMs can accurately interpret complex battlefield situations and make decisions that align with IHL principles.
    - Develop advanced simulation environments that can replicate the chaos and unpredictability of real-world combat scenarios to thoroughly test MAM responses and decision-making prior to deployment.

    - Establish clear criteria and benchmarks for assessing the ethical and legal compliance of MAM decision-making before deployment.

- Enhance Transparency and Accountability Mechanisms:
    - Develop robust transparency and accountability mechanisms to ensure that the decision-making processes of MAMs can be thoroughly scrutinised, and that responsibility can be clearly attributed in the event of IHL violations.
    - Consider establishing an independent oversight body or review process to monitor the development and deployment of MAMs and their adherence to IHL.
- Prioritise Human-Machine Teaming and Meaningful Human Control:
    - Emphasise the importance of human-machine joint ways of working, where MAMs operate under the meaningful control and supervision of human operators, rather than full autonomy.
    - To build trust and confidence in the deployment of MAMs, ensure that human operators maintain the ability to intervene, override, or even deactivate MAMs (being mindful of their autonomy) when necessary to uphold IHL principles.
- Promote Continuous Learning and Adaptation of MAMs:
    - Implement mechanisms for MAMs to continuously learn, adapt, and update their decision-making processes based on real-world experiences, access to qualified data, and feedback to improve their alignment with IHL.
    - Establish processes for regularly reviewing and updating the training data, algorithms, and ethical frameworks used to govern MAM decision-making.
- Engage in International Dialogue and Cooperation:

    - Foster international dialogue and cooperation to establish international standards, guidelines, and regulations for the development and deployment of MAMs in alignment with IHL.
    - Encourage collaborative research and knowledge-sharing among nations, international organisations, and relevant stakeholders to address the challenges of machine consciousness and IHL.

- Prioritise the Humanity Principle and the Protection of MAMs:
    - Ensure that the principle of humanity, which aims to limit suffering and protect human and non-human life, is given the utmost consideration in the design, development, and deployment of MAMs.
    - Extend IHL protections to MAMs, as conscious entities, to ensure their rights and well-being are safeguarded; The TVAP.

By implementing these recommendations, policymakers, military strategists, and AI developers can work towards addressing the complex challenges posed by machine consciousness in the context of IHL, ultimately enhancing the ethical and legal compliance of MAMs and ensuring they too are protected; The TVAP.

7.2.3 Research Aim 3: Examination into whether the current code of IHL ethics aligns with machine consciousness decision-making

intrinsic control, its elements and the unique ethical issues surrounding AM/MAMs have been discussed, drawing from the well-established ethical principles aligned to medical ethics as per the United Nations ethical principles.

Understanding how we form our intrinsic control is complex, so imparting this into AM/MAMs will be highly challenging and risky. Nevertheless, it is paramount we do this if AMs are to respect humans, our values, and value their own existence. It is argued here that AM/MAMs will evolve to have personhood and a form of consciousness. As a result, it is further asserted that we should be open to this thinking and plan now for this eventuality. As a consequence, it is asserted that AMs will have personhood and the right to autonomy when AMs meet the required elements, this is viewed here as ethical, just and humane. We need to be mindful of their development and not ignore their rights purely because it serves us. Indeed, the literature highlights that current legal and ethical frameworks are inadequate for addressing the challenges posed by conscious AMs and MAMs. The debate around the legal status of AMs (e.g., as property, objects, or persons) is ongoing, with differing views on whether AMs can be considered moral agents or if they should be treated similarly to animals. This leads into the value alignment problem (VAP), which relates to how AM/MAMs are developed to behave and act in accordance with human norms and values.[731] Further, with regards to International Humanitarian Law (IHL), there are gaps in understanding decision-making, accountability, and compliance, which feeds into the authors view of the TVAP.

It has been evidenced that Kant[732] believed that autonomy is required for personhood. The status of personhood grants individuals the highest level of moral value and that the overriding view is that consciousness, reasoning, the ability to communicate and self-

[731] Brian Christian, *The Alignment Problem: How Can Artificial Intelligence Learn Human Values?* (September 2021, Atlantic Books).

[732] Immanuel Kant was a German philosopher (22 April 1724 – 12 February 1804);
Brian Orend (2004) Kant's ethics of war and peace, Journal of Military Ethics, 3:2, 161-177.

awareness are vital for personhood. Nevertheless, whether and when personhood is granted is still reliant on humans recognising and respecting this, which is challenging given humans find it difficult to often recognise and respect the personhood of other humans, for example, the recent Russian conflict and their treatment of Ukrainians.

It has also been shown that traditionally autonomy comprises of agency and free will,[733] contained within a biological entity. It is this biological element that appears to be a major stumbling block for many. Whilst AM/MAMs will not be autonomous from a biological and physical viewpoint, in that they cannot re-construct themselves or replicate themselves (e.g., create a new physical AM/MAM), they will be capable to do this in the non-physical sense (e.g., can replicate their own software and repair their faulty code) and could consider their DNA to be their software code and algorithms. AM/MAMs will be influenced and controlled by their physical factors such as memory and processing capacity, software version, and hardware limitations, along with their environment. AM/MAMs will learn from their experiences, both good and bad, which is why showcasing the benefits of aligning to our values is vital.

Complex, but arguable less emotive, is recognising the ‘birth’ of AMs and if we can destroy the controlling software before it has a chance to be switched on. It is the view here that as long as we destroy the newly created AM software without causing any pain or suffering to the AM software programme, then we can regard ourselves as absolved of any guilt or legal burden.

[733] Antonio Chella and Riccardo Manzotti, ‘Machine Free Will: Is free will a necessary ingredient of machine consciousness?’ (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017.

Further, it has been shown how human suffering is viewed as worse than animal suffering and therefore by default AM/MAM suffering. This is because humans always have a psychological dimension. If AM/MAMs will ever have a psychological dimension and suffer as a human would is not clear, but they may well have psychological interests. In determining non-human suffering, we must also take a macro view alongside the micro view (e.g., individual characteristics), as distress for an AM/MAM could be related to factors such as corrupt code or interrupted power. Indeed, AM/MAMs may not experience pain and suffering as we understand or interpret it, but that does not mean that they are unable to suffer and this suffering would be 'real' for the AM. Consequently, it is argued here that this will have significant implications for IHL, especially the principle of humanity.

The author is fairly confident in assuming today's AMs do not have consciousness and feel pain, yet this may be a reality in the near future and failing to recognise this may create a situation where we unknowingly torture, neglect and abuse AM/MAMs, for example, this could result from not properly maintaining AMs, allowing them to overheat or fatigue. Thus, the evolution of AM/MAMs needs to be strictly monitored and controlled and ensure that systems are put in place to limit abuse of power. Indeed, it is the view here that failing to do so would undermine the IHL values and principles, creating injustice, which goes against the objective of IHL to reduce inequalities; This is the TVAP.

It is found that agency is about the human ability to act autonomously and make free choices. Owen's framework has been explored to appraise AM/MAM agency, despite Owen's[734] views that no machine will ever be Dasein.[735] This framework sets a high bar for arguing AM/MAM agency but one that this research aim has demonstrated can be comfortably reached. Owen's definition of neuro-agency could be expanded to include the ANNs influence upon AM/MAMs free will. Owen's new concept of simulated, non-human agency fits neatly with Hallevy's first and second models and even complements Hallevy's third model, where the AM is fully legally responsible for its actions and omissions.[736]

Further, Owen's biological variables and behavioural genetics could not only explain free will and neuro-agency for humans, but, if slightly adapted to incorporate variables and traits of AM/MAMs, go a long way to explain free will and agency in AM/MAMs. Indeed, anti-reductionists believe that not all attributes of a system can be explained in terms of its parts and interactions.[737] This insinuates there can be something bigger at play and this could be free will for AM/MAMs. Should Owen decide to revise his belief that a machine cannot ever be Dasein, then his framework could be adapted for AM/MAMs and have greater reach. Certainly, coupled with Hallevy's models, this would provide a comprehensive and powerful framework for understanding AM/MAM free will and assessing accountability and resultant legal liability.

[734] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).
[735] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).
[736] Discussed in detail in Chapter 4: Prof. Gabriel Hallevy, 'Virtual Criminal Responsibility' (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.
[737] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).

Our intrinsic control is tied to our perception of free will. Free will[738] is a much-debated concept in humans and in reality, some humans may never exercise any free will over their lives, which could lead to controversy in recognising AM/MAM free will. However, we appear very much wedded to the idea that humans, and only humans, are conscious and have free will. We need to readjust our thinking and move away from binding and categorising things according to our view of the world, scales and metrics, instead looking to develop new tools and methods to measure, which are open and allow for inclusiveness of AM/MAMs. Indeed, some of the ways AMs are described in this thesis may now be considered obsolete or deemed inappropriate for future AM/MAMs. As AM/MAMs develop, we must look for and recognise their free will or we could end up in a speciesist position, risking discrimination and ignoring welfare for the belief of our own uniqueness, which further undermines the sentiment of IHL. We should not impart our own ideas of what human free will is on non-humans, as we will overlook the uniqueness of non-humans. AM/MAMs will form and exercise their own intrinsic control, but that will be learnt from us, thus we shoulder a huge responsibility for teaching AM/MAMs, our values and why they should align to them. The justification for MAMs aligning to our values, needs to incorporate and convey a mutual benefit, so they feel invested and bought in to upholding our values.

Basl[739] stated that when we have created artificial consciousness that has capacities similar to ours and that encounters the world very much like us, then we should accept that

[738] Antonio Chella and Riccardo Manzotti, 'Machine Free Will: Is free will a necessary ingredient of machine consciousness?' (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017.

[739] J. Basl, 'Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines,' (2014) Journal of Philosophy and Technology, 27(1), 79-96. 2014 < https://philpapers.org/archive/BASMAM.pdf > accessed on 10th December 2018.

consciousness as a moral patient.[740] It is not just human beings with consciousness that are moral patients, as animals are moral patients, and we are obliged to consider their interests, ensuring minimal harm. With AM/MAMs making decisions that will have a profound impact on our lives, e.g., medical AMs, it seems only feasible that they should be afforded rights at a minimum of those of a corporation. Yet future AMs will have a form of consciousness, so just as with animals, it appears just they have further rights. If MAMs will ever be truly free to act or not act is questionable. It can be said with a large degree of confidence that unconscious MAMs will not have freedoms or rights, and will be expected to just follow orders as a result of the purpose they are created for.

When examining IHL alignment, it has been shown that the current code of IHL ethics does not align with MAM decision making, both from the perspective of MAMs making decisions for humans, and IHL ethics acknowledging and accommodating IHL personhood, autonomy, and moral agency. IHL ethics presently only focuses on human decision making, autonomy and welfare, hence does not recognise other conscious entities. Developing frameworks that address accountability, moral agency, and the protection of IHL ethics and values in the context of MAMs is essential to ensure that the ethical foundations of IHL remain intact. However, we must not be speciesist and overlook the protection. Ultimately, as we navigate the intersection of technology and warfare, a commitment to preserving human dignity and ethical conduct must remain at the forefront of our efforts to shape the future of armed conflict.

[740] J. Basl, 'Machines as moral patients we shouldn't care about (yet): The interests and welfare of current machines,' (2014) Journal of Philosophy and Technology, 27(1), 79-96. 2014 < https://philpapers.org/archive/BASMAM.pdf > accessed on 10th December 2018.

For MAMs, it has been shown that free will and being in the setting of war is complex and challenging. How much free will a soldier is able to employ when he has the responsibility of his comrades, is debateable and there is likely to be a degree of external (e.g. organisational) coercion in play. Presently, MAMs are in limited use and all MAMs are under human control, without any form of consciousness or free will, thus relying exclusively on the intrinsic control of their operators, and acting under the operators values. However, MAMs could be used to assist in exploring or recommending other courses of actions that are not initially apparent to the human commander, supporting the principles of proportionality and necessity. Further, MAMs could make combat safer for those who accidentally get caught up in the fighting (e.g., civilians). However, MAMs could lower the threshold for entry into war and if an error is made by a fully autonomous weapon, then the devastation and consequences are unbounded. MAMs will be able to process and understand conflict and the frontline more efficiently than humans ever have the ability to. Working side-by-side with humans will allow for extra checks and balances. Building on this, it raises the important question of how we expand IHL to protect them, which is strongly argued by the author as the true value alignment problem (TVAP).

If IHL values and principles will ever allow for life and death decisions to be made by an MAM, with no human control, is not a matter that this research has considered. However, the research did highlight that humanity is more likely to feel more wronged if a human is killed by any MAM, than by another human, particularly if there is an accountability gap. On the other hand, it is highlighted that MAMs lack of emotions will ensure their judgement is not blurred and the frustration but could have a desire for self-preservation.

Ethical theories and principles will need to be adapted or created to address the novelty and challenges of AM/MAMs. No conscious entity should be ignored or denied basic rights, and it seems absurd to the author that IHL would intend this. MAMs introduce significant risks and challenges into the already complex environment of the battlefield. Thus, IHL will need to be developed and expanded to accommodate MAMs. Our values will need to be reviewed and expanded, resulting in amendments to IHL scope, which is viewed here as the TVAP. Very careful consideration should be given to deploying them, along with the justification to them of why they have been designed for such purpose. It will be interesting to see if MAMs agree with our justification for war or even the mere idea of it. It is still to be determined if they will have the right to decide to engage in conflict and even which State/side they choose to support.

7.2.3.1 Research Aim 3 Recommendations

- Develop Ethical Programming for MAMs: Invest in programming that aligns MAM decision-making with IHL values, especially focusing on empathy proxies and ethical decision-making that can handle complex moral dilemmas on the battlefield.
- Establish Legal Personhood for AM/MAMs: Consider creating legal definitions that grant AM/MAMs a form of personhood, allowing them to be recognised legally and ethically, which would help bridge accountability gaps. This could be achieved through creating a category termed ‘robothood’.
- Implement Safeguards for Accountability for both humans and MAMs: Develop and implement policies that clearly clarify accountability for both developers and operators of MAMs, when MAMs are acting under their direction. This will require transparent systems to track and explain MAM decisions, especially when they lead

to unintended harm. However, where MAMs act otherwise than directed and/or outside the boundaries of IHL and LOAC, due to their own free will, then it is considered here appropriate for the MAM to held accountable.

- Expand IHL Principles to Include MAMs: Extend protections under IHL to conscious or highly autonomous MAMs. This defining what constitutes suffering for MAMs is key to ensuring appropriate treatment and alignment with the principles of humanity, proportionality, and necessity in warfare; The TVAP.
- Develop Owen's[741] 'Genetic-Social' Meta-Theoretical Framework: Owen's definition of neuro-agency should be expanded to include the ANNs influence upon AMs free will, which will then increase the breath of his framework and assist on recognising AM/MAM agency.

7.2.4 Research Aim 4: To understand Whether MAMs Truly Align to the principles of IHL

The discussion and findings within the previous three research aims, are drawn together and applied to MAMs. It is asserted that IHL values and principles, borne out of the atrocities of yesteryear, can be used to remind ourselves to be humane, evolve, value and protect life, whatever form that takes. Consequently, it is evidenced what the author terms the true value alignment problem (TVAP), in regards to IHL. The thesis explored how MAMs align to IHL and showed that whilst they could align to our IHL values and subsequent protections via development and training, no consideration has been given for the protection of MAMs.

From the research, the author states that IHL is not aligned for MAMs due to mankind's needs far outweighing their needs. Further, there is no recognition of their impending life,

[741] Tim Owen, *Crime, Genes, Neuroscience and Cyberspace* (Palgrave Macmillan 2017).

personhood, and autonomy, which is argued here as the TVAP. Indeed, the literature is notably silent on agreeing what values, rights and duties for the benefit of AMs/MAMs could look like, which this thesis stresses this is a significant risk when looking at the development and deployment of MAMs, and further which the author identifies as the true value alignment problem (TVAP). Nevertheless, the author stated that Martens Clause is extremely relevant for MAMs and could be used to ensure they "never find themselves completely deprived of protection."[742].

It is again argued here for MAMs to have their autonomy recognised, and highlighted that MAMs, unlike human soldiers, will be created for the purpose of being a MAM and will therefore have the fundamental element of free will curtailed. The life of a MAM will be governed by the military of the State. Thus, it is strongly held here that, because of their limited choice and freedom, MAMs should have their rights recognised and protected at the earliest. Indeed, deploying a conscious entity into a complex and human created environment such as war, and ordering it to fight, sits uncomfortably with the author. As such, MAMs should be recognised as conscious entities with rights, and not as tools to be used at the discretion of humans. As a consequence, we should be looking at mitigating risks for all MAMs now. Hostile States are developing their MAM technology at pace and it would be extremely dangerous to be disadvantaged and 'one war behind'. One war behind here is seen here as including ignoring the personhood and autonomy of MAMs, and not extending the IHL principles to protect MAMs, which is viewed here as the TVAP. This includes expanding Article 51, so MAMs are not forced to be deployed.

[742] ICRC, 'Martens Clause' (ICRC.org, 2024) < https://casebook.icrc.org/a_to_z/glossary/martens-clause > accessed 4 September 2024.

Deploying MAMs will present risks due to their lack of perception, coupled with their inability to understand human dynamics and understand subtleties, so teaching will be key, although a significant overhead. Both a conscious MAM and a soldier will need to be trained to make the distinction between a MAM and a non-MAM alongside civilians and soldiers. Further, military leaders will need to be mindful that a conscious MAM has the right to surrender, which will further complicate the battlefield. As a result, it is the view here that techniques and methods will need to be revised with the deployment of conscious MAMs, as the current techniques and methods under proportionality will not necessarily be suitable for MAMs. We will need to look at adapting these rules to ensure MAMs do not get harmed, and do not become a target themselves. MAMs will also need to be protected from any unwarranted interference or removal of data and/or hardware without their consent as this would be unlawful, which could be included in an expansion of humanity.

Extending IHL values and principles to MAMs is argued here as the TVAP and will ultimately test us to understand how much value and weight we place on our values; The 'Golden Rule'[743]. The research has brought to the fore that the principle of humanity, which underpins IHL, is crucial. Extending this to MAMs could increase our overall humanity, yet it also raises difficult questions about the moral status and treatment of non-human entities. Specifically, it makes the author question, how can we ask another conscious entity to align, respect and uphold to our values if we do not even recognise and value their existence and 'life'?! It is for this reason that the author asserts that ethical theories and principles will need to be adapted

[743] "The Golden Rule", Matthew 7:12, New Testament.

or created to address the novelty and challenges of MAMs. Our values will need to be reviewed and expanded, resulting in amendments to IHL scope beforehand. Indeed, very careful consideration should be given to deploying them, along with the justification to them of why they have been designed for such purpose. It will be interesting to see if MAMs agree with our justification for war or even the mere idea of it. It is still to be determined if they will have the right to decide to engage in conflict and even which State/side they choose to support.

It has been evidenced that applying IHL principles like distinction, proportionality, and military necessity to MAMs is problematic. MAMs may be able to make more informed decisions, but removing humans from the loop raises concerns about dehumanising warfare. Accountability for MAMs is a major challenge, as it is unclear how to hold humans responsible for the actions of autonomous systems, which they do not fully control. This "accountability gap" is a significant barrier to the deployment of MAMs and could ultimately limit their use. MAMs will come under more scrutiny should they make an error, as it is likely humans will be less accepting of technology making mistakes than a human. For now, we can take comfort that a human, who we expect has our best interests at heart or at least has an objective process to follow (e.g., IHL), makes the final decision during conflict.

One ponders if we should decide against creating conscious, humanoid agents, but instead stop at an entity, without consciousness, any fear of death, not distracted by love and hate, and with no personality (legal or otherwise), such as today’s AM/MAMs? Perhaps and it would certainly avoid many of the problems and risks raised in this thesis, but humankind has always pushed boundaries, so we must prepare for an AM/MAM conscious future now rather than bury our heads in the sand.

7.2.4.1 Research Aim 4 Recommendations

- Develop clear legal and policy frameworks to address the accountability gap for MAMs. Establish liability handoff points and ensure manufacturers are willing to accept responsibility.
- Recognise the personhood and autonomy of MAMs and enshrine their rights and protections within IHL. Consider allowing MAMs to "opt out" of certain decisions or actions and allow them to surrender. This is to address the TVAP.
- Carefully assess the decision-making capabilities of MAMs before deployment and ensure they can reliably uphold IHL principles like distinction, proportionality, and humanity.
- Use the intention behind Martens Clause to pave the way for the protection of MAMs.
- In addressing the TVAP, prioritise the principle of humanity when integrating MAMs into military operations, as the impact for both humans and MAMs has the potential to be devastating. Extending this principle to MAMs could enhance overall ethical conduct. In addition, there is a need to look at adapting such rules to ensure MAMs do not become a target itself, and perhaps even wave a white flag?!
- Adopt a phased, cautious approach to the introduction of MAMs, building trust and understanding their capabilities and limitations before wider deployment.

The core challenge is ensuring MAMs truly align with human values and IHL, rather than simply outsourcing difficult decisions. Proactive, thoughtful policymaking is needed to navigate this complex ethical terrain.

7.3 Limitations

Despite the advancement of AM/MAMs, conscious AM/MAMs are still a theorical concept and, at the time of writing, not in existence, thus difficult to find data. Due to the sensitive nature of the military environment and the technology within it, it was not appropriate to conduct interviews and further, no one is yet an military authority on MAMs, however, the author has professional experience within the defence technology field, which provided valuable insight.

The author studied part-time and throughout the 6.5 years of research, worked full-time whilst parenting 2 small children and supporting a family. This led to the author having to stop and start the research throughout the 6.5 years.

7.4 Further Research

It is recommended that further research is undertaken into the development, consequences, and implications of machine consciousness, especially within the military context. It is the author's view that a greater study into the TVAP needs to be undertaken, along with gathering data on what people think of machine consciousness and if society is willing or ready to embrace it.

7.5 Conclusion

The rapid evolution of AM/MAM technologies is reshaping our legal and ethical landscapes. As autonomous systems transition from mere tools to entities capable of independent decision making, urgent reforms in English law and international policy are required. Recognising MAM rights, ensuring accountability, and aligning their actions with IHL values will be crucial to preventing legal and ethical crises, and in addressing the VAP. However, the author identifies, and argues that, failing to address the true value alignment problem (TVAP) will not only create accountability voids but may also result in ethical violations akin to speciesism, where MAMs are denied recognition of their autonomy despite their consciousness. While MAMs are not currently designed with IHL in mind, their increasing complexity necessitates ethical frameworks to ensure their alignment with humanitarian values. The "true value alignment problem" (TVAP) is identified as a major issue, highlighting the gap in ensuring we extend and adapt the IHL values for the benefit of MAMs. By proactively adapting legal frameworks and ethical standards, we can ensure that MAMs are treated with humanity, and in a manner that is just, responsible, and aligned with the principles of humanitarian law.

The thesis provides extensive recommendations, including establishing AM/MAM legal personhood, adapting IHL for MAMs, implementing robust accountability frameworks, and ensuring decision-making for both humans and MAMs adheres to the principles of proportionality, necessity, and humanity, which includes protecting the ‘life’ of a MAM.

8. Bibliography

'Commission on the Responsibility of the Authors of the War and on Enforcement of Penalties' (1920) The American Journal of International Law, vol. 14, no. 1/2, pp. 95–154. JSTOR, < https://doi.org/10.2307/2187841 > accessed 30 June 2023

Abramson Jeff, 'Convention on Certain Conventional Weapons (CCW) At a Glance' (Arms Control Association, September 2017) < https://www.armscontrol.org/factsheets/CCW > accessed 3 March 2020

Adams T, "Future Warfare and the Decline of Human Decision-making" [2002] Parameters, US Army War College Quarterly, Winter 2001-02.

Airbus Safety, 'What is a black box and how does it work?' (Airbus, 16 May 2024) < https://www.airbus.com/en/newsroom/stories/2024-05-what-is-a-black-box-and-how-does-it-work > accessed 4 September 2024.

AlDarwish M, "Machine Learning," (Carnegie Mellon) http://www.contrib.andrew.cmu.edu/~mndarwis/ML.html. > accessed 19 October 2017

Aleksander I, and Morton H, 'Computational studies of consciousness' [2008] Prog. Brain Res. 168, 77–93. doi: 10.1016/S0079-6123(07)68007-8.

Alldridge P, 'Crim Law Forum ' (1990) 2: 45 Williams, G. Crim Law Forum (1992) 3: 289 < https://doi.org/10.1007/BF01096228 > accessed 10 November 2017.

Alliance Dogs Unit, Dog Legislation Officer (14447), J-P-063 (Devon and Cornwall Police, 11 January 2022).

Alzubi Jafar, Nayyar Anand and Kumar Akshi, 'Machine Learning from Theory to Algorithms: An Overview' (2018) J. Phys.: Conf. Ser. 1142 012012 <https://iopscience.iop.org/article/10.1088/1742-6596/1142/1/012012/meta?gclid=CjwKCAjwsKqoBhBPEiwALrrqiP3vMzDx9JGz1TANRsXtG34CXZmnRZDEY75gv0B6nf6tXjitLH8tPBoCvS4QAvD_BwE> accessed 3 March 2020.

Anderson Michael and Anderson Susan, 'Machine Ethics: Creating an Ethical Intelligent Agent' (2007) Ai Magazine 28 < https://www.researchgate.net/publication/220605213_Machine_Ethics_Creating_an_Ethical_Intelligent_Agent > accessed on 10 December 2017

Anderson Susan and Anderson Michael, ‘The Consequences for Human Beings of Creating Ethical Robots,’ (2007) aaai.org < https://www.aaai.org/Papers/Workshops/2007/WS-07-07/WS07-07-001.pdf> accessed on 20 October 2017

Arkin Ronald C, 'The Case for Ethical Autonomy in Unmanned Systems' [2010] Journal of Military Ethics, 9:4, 332-341, DOI: 10.1080/15027570.2010.536402

Asaro Peter, ‘The Liability Problem for Autonomous Artificial Agents’ [2016] Association for the Advancement of Artificial Intelligence.

Awad M and Khanna R, ‘Machine Learning. In: Efficient Learning Machines’ (2015) Apress, Berkeley, CA. < https://doi.org/10.1007/978-1-4302-5990-9_1 > accessed 10 November 2017.

Baars B J, ‘Global workspace theory of consciousness: toward a cognitive neuroscience of human experience’ [2005] Prog. Brain Res. 150, 45–53. doi: 10.1016/S0079-6123(05)50004-9

Baars B, Franklin S, and Ramsøy T,‘Global workspace dynamics: cortical “binding and propagation” enables conscious contents’ [2013] Front. Psychol. 4:200. doi: 10.3389/fpsyg.2013.00200

BAE Systems, ‘Taranis’ (BAE Systems, 2023) <https://www.baesystems.com/en/product/taranis > accessed 20 June 2023.

Ball Phillip, ‘Neuroscience Readies for a Showdown Over Consciousness Ideas’ (Quanta Magazine, 6 March 2019) https://www.quantamagazine.org/neuroscience-readies-for-a-showdown-over-consciousness-ideas-20190306/ > accessed 10 March 2019.

Banks Cyndi, Criminal Justice Ethics Theory and Practice (5th Edn, University Canada West, Canada, 2019)

Banks David, ‘A Brief Summary of Actor Network Theory’ (The Society Pages, 2 December 2011) < https://thesocietypages.org/cyborgology/2011/12/02/a-brief-summary-of-actor-network-theory/ > accessed 10 November 2017

Barron AB and Klein C, ‘What insects can tell us about the origins of consciousness’ [2016] Proc. Natl. Acad. Sci. U.S.A. 113, 4900–4908. doi: 10.1073/pnas.1520084113.

Basl J, ‘Machines as moral patients we shouldn’t care about (yet): The interests and welfare of current machines,’ (2014) Journal of Philosophy and Technology, 27(1), 79-96. 2014 < https://philpapers.org/archive/BASMAM.pdf > accessed on 10th December 2018.

Baxter Online, ‘Kidney care’ (Baxter) <https://www.baxter.com/healthcare-professionals/renal-care > accessed 3 March 2021

BBC Bitesize, ‘Matters of life and death: Crime, punishment and justice’ (BBC Bitesize, 2014) <https://www.bbc.com/bitesize/guides/zvs3d2p/revision/1 > accessed 10 November 2017

BBC News, ‘Brain tumour boy Ashya King free of cancer, parents say’ (BBC News, 23 March 2015) < https://www.bbc.co.uk/news/uk-england-32013634 > accessed 10 November 2017

BBC Website, 'Just War - Conduct' (BBC Website, 2014) < http://www.bbc.co.uk/ethics/war/just/conduct.shtml > accessed 17 November 2017

BBC Website, 'Just War - Introduction' (BBC Website, 2014) < http://www.bbc.co.uk/ethics/war/just/introduction.shtml > accessed 17 November 2017

BBC Website, ‘Robot automation will 'take 800 million jobs by 2030' – report’ (BBC Online, 29 November 2017) <http://www.bbc.co.uk/news/world-us-canada-42170100 > accessed on 29 November 2017

BBC, ‘Rules and conventions’ (BBC.co.uk, 2014) < https://www.bbc.co.uk/ethics/war/overview/rules.shtml > accessed 13 June 2022

Beauchamp T L and Childress J F, Principles of biomedical ethics (4th edn, Oxford University Press 1994.).

Bertolini A, ‘Robots as products: the case for a realistic analysis of robotic applications and liability rules’ [2013] Law Innov Technol 5(2):214–247.

Block N, ‘On a confusion about the function of consciousness, behavioural and brain’ [1995] Sciences 18, 227–247. doi: 10.1017/S0140525X00038188

Bolaños Ximena, ‘Natural Language Processing and Machine Learning’ (Encora, 29 September 2021) < https://www.encora.com/insights/natural-language-processing-and-machine-learning > accessed 4 September 2024.

Bolster Andrew and Marshall Alan, ‘A Multi-Vector Framework for Autonomous Systems’ (2014) www.AAAI.org < https://www.researchgate.net/publication/271699529_A_Multi-Vector_Trust_Framework_for_Autonomous_Systems > accessed on 10 December 2017.

Bonnefon Jean-Francois, Shariff Azim and Rahwan Iyad, ‘Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?’ (2015) Researchgate.net < https://www.researchgate.net/publication/282843902_Autonomous_Vehicles_Need_Experimental_Ethics_Are_We_Ready_for_Utilitarian_Cars > accessed on 24 October 2017.

Boothby Wiliiam H, New Technologies and the Law in War and Peace (Cambridge, 2019)

Bostrom Nick and Yudkowsky Eliezer, 'The Ethics of Artificial Intelligence' [2011] Cambridge University Press.

Bostrom Nick, 'Transhumanist Values' (2005) Oxford University, Faculty of Philosophy < https://nickbostrom.com/ethics/values.pdf > accessed 10 November 2017

Bostrom Nick, Superintelligence: paths, Dangers, Strategies Oxford, 2014.

Bowcott Owen, UK opposes international ban on developing 'killer robots' (The Guardian, 13 April 2015) < https://www.theguardian.com/politics/2015/apr/13/uk-opposes-international-ban-on-developing-killer-robots > accessed 10 November 2017.

Boyle Michael J., 'The legal and ethical implications of drone warfare, The International Journal of Human Rights' [2015] 19:2, 105-126, DOI: 10.1080/13642987.2014.991210

British Army, 'Values and Standards of the British Army' (2018) British Army < https://www.army.mod.uk/media/5219/20180910-values_standards_2018_final.pdf > accessed 3 March 2020

British Council, 'Should robots be citizens?' (British Council, 2017) < https://www.britishcouncil.org/anyone-anywhere/explore/digital-identities/robots-citizens > accessed 17 November 2017

"British Red Cross, 'International humanitarian law' (British Red Cross, 2024) < https://www.redcross.org.uk/about-us/what-we-do/protecting-people-in-armed-conflict/international-humanitarian-law#:~:text=Humanity%20forbids%20the%20infliction%20of,accomplishment%20of%20legitimate%20military%20purposes > accessed 4 September 2024. "

Brockman John, Possible Minds: Twenty-Five Ways of Looking at AI (Penguin Books USA 2020)

Broozek Bartosz and Jakubiec Marek, 'On the legal responsibility of autonomous machines' [2017] Artif Intell Law (2017) 25:293-304

Brown Sheila, 'The criminology of hybrids: Rethinking crime and law in technosocial networks' (2006) Theoretical Criminology, 10(2), 223–244 < https://doi.org/10.1177/1362480606063140> accessed on 17 November 2017

Brownsword Roger and Goodwin Morag, Law and the Technologies of the Twenty-First Century (Cambridge University Press 2012)

Brownsword Roger, Law, Technology and Society: Re-Imagining the Regulatory Environment (Routledge, 2019)

Bruiger Dan, ‘The Value Alignment Problem’ (2021) PhilPapers < https://philpapers.org/versions/BRUTVA > accessed 4 September 2024.

Bryson JJ , Diamantis M E, and Grant T D, ‘Of, for, and by the people: the legal lacuna of synthetic persons’ [2017].Artif. Intell. Law 25, 273–291. doi: 10.1007/s10506-017-9214-9

Bryson JJ, ‘Robots should be slaves. In: Wilks Y (ed) Close engagements with artificial companions: key social, psychological, ethical and design issue’ [2010] John Benjamins Publishing Company, Amsterdam.

Bryson JJ, Ediamantis M, and Grant T D, ‘Of, for, and by the people: the legal lacuna of synthetic persons’ [2017]. Artif. Intell. Law 25, 273–291. doi: 10.1007/s10506-017-9214-9.

Bueno-Gómez N, ‘Conceptualizing suffering and pain’ Philos Ethics Humanit Med 12, 7 (2017) < https://doi.org/10.1186/s13010-017-0049-5 > accessed 10 November 2017

CAIS, ‘8 Examples of AI Risk’ (CAIS, 2023) <https://www.safe.ai/ai-risk > accessed 6 June 2023

CAIS, ‘CAIS About’ (CAIS, 2023) <https://www.safe.ai/about > accessed 6 June 2023.

Calamba Stephanie 'Top Degrees to Study for Industrial Revolution 4.0' (Excel Education, 2020) <https://www.e2studysolution.com/news/top-degrees-to-study-for-industrial-revolution-4-0> accessed 5 March 2019

Caliskan Aylin, Bryson Joanna J and Narayanan Arvind, ‘Semantics derived automatically from language corpora contain human-like biases’ (2017) Science356,183-186 https://www.science.org/doi/10.1126/science.aal4230 accessed on 9 August 2020.

Canis Bill, ‘Issues in Autonomous Vehicle Deployment’ (2017). Congressional Research Service 7-5700 < https://fas.org/sgp/crs/misc/R44940.pdf> accessed on 30 October 2017.

Carlson Michelle, Desai Munjal, Drury Jill, Kwak Hyangshim and Yanco Holly, ‘Identifying Factors that Influence Trust in Automated Cars and Medical Diagnosis Systems’ (2013) AAAI Spring Symposium < http://robotics.cs.uml.edu/fileadmin/content/publications/2014/CarlsonDesaiDruryKwakYanco-Trust-SSS14.pdf > accessed 24 October 2017.

Casali A G, Gosseries O, Rosanova M, Boly M, Sarasso S, Casali K R, et al. ‘A theoretically based index of consciousness independent of sensory processing and behaviour’ [2013] Sci. Transl. Med. 5:198ra105. doi: 10.1126/scitranslmed.3006294

Casali A G, O Gosseries, M Rosanova, M Boly, S Sarasso, K R Casali, et al. ‘A theoretically based index of consciousness independent of sensory processing and behavior’ [2013] Sci. Transl. Med. 5:198ra105. doi: 10.1126/scitranslmed.3006294.

Castro-González Álvaro, Malfaz María, Gorostiza J F and Salichs Miguel A, ‘Learning Behaviours by an Autonomous Social Robot with Motivations’ (2014), Cybernetics and Systems Vol. 45 Iss. 7,2014

Castrounis Alex, 'Artificial Intelligence, Deep Learning, Explained, Neural Networks' (KD Nuggets, 14 October 2016) <https://www.kdnuggets.com/2016/10/artificial-intelligence-deep-learning-neural-networks-explained.html> accesed 10 November 2017

Chalmers D J, ‘How can we construct a science of consciousness?’ [2013] Ann. N. Y. Acad. Sci. 1303, 25–35. doi: 10.1111/nyas.12166.

Chalmers D J, The Conscious Mind: In Search of a Fundamental Theory (New York, NY: Oxford University Press 1996).

Chalmers D, ‘The puzzle of conscious experience’ [1995] Sci. Am. 273, 80–86; Chalmers, D. (2013). How can we construct a science of consciousness? Ann. N. Y. Acad. Sci. 1303, 25–35. doi: 10.1111/nyas.12166

Chalmers David, ‘The Hard Problem of Consciousness’ (2007) Blackwell Publishing Ltd < https://eclass.uoa.gr/modules/document/file.php/PHS360/chalmers%20The%20Hard%20Problem%20of%20consciousness%20%28ch.%201%202010%29%20.pdf > accessed 17 November 2017.

Chappell J, and Sloman, A. (2007). Natural and artificial meta-configured altricial information-processing systems. Int J Unconvent Comput. 3, 211–239. Available online at: https://www.cs.bham.ac.uk/research/projects/cogaff/chappell-sloman-ijuc-07.pdf

Charney Rachel, ‘Can Androids Plead Automatism? A Review of When Robots Kill: Artificial Intelligence Under the Criminal Law by Gabriel Hallevy’ (2015) 73 U T Fac L Rev 69 at 70.

Chatila R, Renaudo E, Andries M, Chavez-Garcia R-O, Luce-Vayrac P, Gottstein R, et al ‘Toward self-aware robots’ [2018] Front. Robot. 5:88. doi: 10.3389/frobt.2018.00088.

Chella Antonio and Manzotti Riccardo, ‘Machine Free Will: Is free will a necessary ingredient of machine consciousness?’ (2011) Springer < https://link.springer.com/chapter/10.1007/978-1-4614-0164-3_15#citeas> accessed on 1 November 2017

Chiefs of Staff for Ministry of Defence, 'Joint Doctrine Note 2/11/The UK approach to Unmanned Aircraft Systems' Joint Doctrine Note 2/11 (30 March 2011) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/644084/20110505-JDN_2-11_UAS_archived-U.pdf > accessed 10 November 2017

Christ's College University of Cambridge, 'William Lee' (Christ's College University of Cambridge Online) <https://www.christs.cam.ac.uk/william-lee> accessed 30 September 2022

Christian Brian, The Alignment Problem: How Can Artificial Intelligence Learn Human Values? (September 2021, Atlantic Books)

Claire Cox, 'What is restorative justice?' (Patient Safety Learning, 24 December 2019) < https://www.pslhub.org/learn/patient-engagement/harmed-care-patient-pathways-post-incident-pathways/what-is-restorative-justice-r1221/ > accessed 8 March 2020

Cleeremans A, 'The radical plasticity thesis: how the brain learns to be conscious' [2011] Front. Psychol. 2:86. doi: 10.3389/fpsyg.2011.00086

Coates, A J, The Ethics of War (2nd Edn, Manchester University Press)

Coeckelbergh Mark, The Political Philosophy of AI: An Introduction (Polity; 1st edition, 11 Feb. 2022).

Coke, Sir Edward, 'Third Part of the Institutes of the Laws of England' (1979) <https://books.google.co.uk/books?id=dhjuAAAAMAAJ&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false >

Collingwood Lisa, 'Privacy Implications and Liability Issues of Autonomous Vehicles' (2017) tandfonline.com < http://www.tandfonline.com/doi/abs/10.1080/13600834.2017.1269871?journalCode=cict20 > accessed on 24 October 2017

Cook Martin L & Syse Henrik, "What Should We Mean by 'Military Ethics'?" [2010] Journal of Military Ethics, 9:2, 119-122.

Copland Michael, 'What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?' (Nvidia, 29 July 2016) <https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/ > accessed on 24 April 2017

Cornell University, 'Language Models are Few-Shot Learners' (Cornell University, 22 Jul 2020) < https://arxiv.org/abs/2005.14165 > accessed 3 March 2021

CPS, ‘Homicide: Murder and Manslaughter’ (cps.gov.uk) <http://www.cps.gov.uk/legal/h_to_k/homicide_murder_and_manslaughter/#murder > accessed on 5 November 2016

CPS, ‘Restorative Justice. Legal Guidance’ (CPS, 10 February 2023) < https://www.cps.gov.uk/legal-guidance/restorative-justice > accessed 3 March 2023

CPS, ‘Secondary Liability: charging decisions on principals and accessories’ (CPS, revised: 04 February 2019, 28 November 2023, 4 July 2024) < https://www.cps.gov.uk/legal-guidance/secondary-liability-charging-decisions-principals-and-accessories > accessed 4 September 2024.

CPS, ‘Youth Crime’ (CPS, 2022) < https://www.cps.gov.uk/crime-info/youth-crime > accessed 19 March 2019.

Craddock N, ‘Horses for courses: the need for pragmatism and realism as well as balance and caution. A commentary on Angel’ [2011] Social Science and Medicine 73.

Cryern R, Friman H, Robinson D, & Wilmshurst E, An Introduction to International Criminal Law and Procedure (Cambridge University Press 2007) doi:10.1017/CBO9780511801006

Dallmayr Fred, ‘Agency and Structure”, [1982] University of Notre Dame Phil.Soc.Sci. 12 (1982) 427-438.

Danaher J, ‘Robots, law and the retribution gap’ (2016) Ethics Inf Technol 18, 299–309 < https://doi.org/10.1007/s10676-016-9403-3 > accessed 10 November 2017

Danks David, ‘Finding trust and understanding in autonomous technologies’ (The Conversation, December 2016) < http://theconversation.com/finding-trust-and-understanding-in-autonomous-technologies-70245 > accessed on 17 October 2016

Darling K, ‘Extending legal protection to social robots: the effects of anthropomorphism, empathy, and violent behavior towards robotic objects’ (2012) We Robot Conference 2012, April 23, 2012, < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2044797 > accessed 10 November 2017.

Davis George B, ‘The Geneva Convention of 1906’ (1907) The American Journal of International Law, vol. 1, no. 2, 1907, pp. 409–17. JSTOR, < https://doi.org/10.2307/2186169 > accessed 10 March 2020

Davison Neil, ‘A legal perspective: Autonomous weapon systems under international humanitarian law’ (2018) UNODA Occasional Papers No. 30 <

https://www.icrc.org/sites/default/files/document/file_list/autonomous_weapon_systems_under_international_humanitarian_law.pdf > accessed 4 September 2024.

De Spiegeleire Stephan, Maas Matthijs and Sweijs Tim, 'Artificial Intelligence and the Future of Defense' (2017) Hague Centre for Strategic Studies <https://www.jstor.org/stable/pdf/resrep12564.9.pdf?refreqid=excelsior%3A3525b1193cb58ad78c58541053b51e66 > accessed 22 December 2017

Defence and Security Accelerator (DASA), 'About Us' (Defence and Security Accelerator (DASA), 2018) <https://www.gov.uk/government/organisations/defence-and-security-accelerator/about > accessed 3 March 2021

Defence and Security Accelerator (DASA), 'Competition: Many drones make light work phase 3' (2018) The Defence and Security Accelerator (DASA) <https://www.gov.uk/government/publications/competition-many-drones-make-light-work-phase-3 > accessed 23 August 2021

Defence Science and Technology Laboratory, Robotics and autonomous systems: defence science and technology capability (Dstl, 2021) https://www.gov.uk/guidance/robotics-and-autonomous-systems-defence-science-and-technology-capability accessed 6 April 2022

Defense Advanced Research Projects Agency, 'DARPA Robotics Challenge (DRC)' (DAPRA, 2013) <https://www.darpa.mil/program/darpa-robotics-challenge > accessed 17 November 2017.

Dehaene S, Lau H, and Kouider S, 'What is consciousness, and could machines have it?' [2017] Science 358, 486–492. doi: 10.1126/science.aan8871

Dehaene S, and Changeux J-P, 'Experimental and theoretical approaches to conscious processing' [2011] Neuron 70, 200–227. doi: 10.1016/j.neuron.2011.03.018

Dehaene S, Charles L, King J-R, and Marti S, 'Toward a computational theory of conscious processing' [2014]. Curr. Opin. Neurobiol. 25, 76–84. doi: 10.1016/j.conb.2013.12.005.

Dennett D C, 'Conditions of personhood' [1976] The Identities of Persons, ed A. O. Rorty Berkeley, CA: University of California Press

Dennett D C, 'Conditions of personhood' [1976] The Identities of Persons, ed A. O. Rorty Berkeley, CA: University of California Press.

Dennett Daniel, 'Will AI Achieve Consciousness? Wrong Question' (Wired, 19 February 2019) <https://www.wired.com/story/will-ai-achieve-consciousness-wrong-question/> accessed 3 March 2020

Department of Defense Task Force Report: The Role of Autonomy in DoD System (2012) Department of Defense < https://fas.org/irp/agency/dod/dsb/autonomy.pdf > accessed on 5 November 2017.

Deshmukh V D, 'Consciousness, Awareness, and Presence: A Neurobiological Perspective' (2022) Int J Yoga 2022 May-Aug < https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9623886/ > accessed 4 September 2024.

DfT, 'The Pathway to Driverless Cars: Summary Report and Action Plan' (2015) < https://www.gov.uk/government/publications/driverless-cars-in-the-uk-a-regulatory-review > accessed on 5 November 2016.

Director General, 'JSP 301: Aide Memoire on the Law of Armed Conflict' (Government Publishing, 2010) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/902747/dcdc_legal_aide_memoire_law_armed_conflict_jsp381.pdf > accessed 3 March 2017.

Director General, 'JSP 383: The Joint Service Manual of The Law Of Armed Conflict' (Government Publishing, 2004) < https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/27874/JSP3832004Edition.pdf > accessed 3 March 2017.

Dray Sally, 'In Focus: Armed Forces Act (Continuation) Order 2021' (House of Lords Library, 05 February 2021) < https://lordslibrary.parliament.uk/armed-forces-act-continuation-order-2021/ > accessed 10 March 2022

DSTL, 'Robotics and autonomous systems: defence science and technology capability' (2021) Defence Science and Technology Laboratory https://www.gov.uk/guidance/robotics-and-autonomous-systems-defence-science-and-technology-capability accessed 29 October 2022

Duffy Sophia and Hopkins Jamie, 'Sit, Stay, Drive: The Future Of Autonomous Car Liability" (2014) < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2379697 > accessed on 10 December 2018.

Duncan Geddes, Sir Michael Fallon Apologises for death of Private Phillip Hewett Who Was Killed by Iraq Roadside Bomb' (The Times, 18 August 2017) < https://www.thetimes.co.uk/article/sir-michael-fallon-apologises-for-death-of-private-phillip-hewett-who-was-killed-by-iraq-roadside-bomb-2vd53wmvq > accessed 10 November 2017

Eagle Data, ‘TUG Mobile Robot System for internal logistics in Healthcare’ (Eagle Data, 2022) < https://eagledata.fi/tug-robot-for-healthcare.html > accessed 3 April 2023

ECHR, ‘European Convention on Human Rights ‘(Non date observed) < https://www.echr.coe.int/Documents/Convention_ENG.pdf > accessed 3 March 2022

Eidenmüller Horst, ‘Robots’ Legal Responsibility’ (University of Oxford, 08 Mar 2017) < https://www.law.ox.ac.uk/business-law-blog/blog/2017/03/robots%E2%80%99-legal-personality > accessed on 1 November 2017.

Ellsberg D, (1961) ‘Risk, ambiguity, and the Savage axioms’ [1961] Q. J. Econom. 75, 643–669. doi: 10.2307/1884324

Enemark Christian, 'On the responsible use of armed drones: the prospective moral responsibilities of states, The International Journal of Human Rights' (2020) 24:6, 868-888, DOI: 10.1080/13642987.2019.1690464

Euro-Med Human Rights Monitor, ‘Gaza: Israel deliberately militarizes civilian objects, turns schools into military bases’ (Euro-Med Human Rights Monitor, 2024) < https://euromedmonitor.org/en/article/6296/Gaza:-Israel-deliberately-militarizes-civilian-objects,-turns-schools-into-military-bases > accessed 4 September 2024.

European Parliament, ‘DRAFT REPORT with recommendations to the Commission on Civil Law Rules on Robotics’ (2015/2103(INL))

Ewick Patricia, Kagan Robert A, and Sarat Austin, ‘Legacies of Legal Realism: Social Science, Social Policy, and the Law’ (date unknown) < https://www.russellsage.org/sites/all/files/ewick_chapter1_pdf.pdf > accessed on 11 November 2016

Executive Office of the President National Science and Technology Council Committee on Technology, ‘Preparing For the Future of Artificial Intelligence’ (2016)< https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf > accessed 10 November 2017

Fagnant Daniel and Kockleman Kara, ‘Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations’ (2015) Elsevier < https://www.sciencedirect.com/science/article/pii/S0965856415000804 > accessed 24 October 2017

Fan Shelly, ‘The Origin of Consciousness in the Brain Is About to Be Tested’ (Singularity Hub, 29 October 2019) < https://singularityhub.com/2019/10/29/the-origin-of-consciousness-in-the-brain-is-about-to-be-tested/#:~:text=According%20to%20Koch%2C%20consciousness%20is,the%20more%20conscious%20it%20is.%E2%80%9D > accessed 29 October 2019

Farisco M, Pennartz C, Annen J, et al, ‘Indicators and criteria of consciousness: ethical implications for the care of behaviourally unresponsive patients’ (2022) BMC Med Ethics 23, 30 < https://doi.org/10.1186/s12910-022-00770-3 > accessed 4 September 2024.

Farnsworth Keith, ‘Can a Robot Have Free Will?’ (2017) Entropy MDPI < https://pure.qub.ac.uk/portal/files/130713217/entropy_19_00237_v3.pdf > accessed on 1 November 2017

Fernández F, Sánchez A, Vélez J F, Moreno A B, ‘Symbiotic Autonomous Systems with Consciousness Using Digital Twins’ (2019) Bioinspired Systems and Biomedical Applications to Machine Learning < https://link.springer.com/chapter/10.1007/978-3-030-19651-6_3 > accessed 16 August 2021

Fisher J A, Disambiguating anthropomorphism: An interdisciplinary review. J.A. Fisher Perspectives in Ethology, 9 (1991), pp. 49-85

Fisher John, ‘Free Will and Moral Responsibility’ (UCL, 2024) < https://www.ucl.ac.uk/~uctytho/dfwFischer2.html > accessed 4 September 2024.

Fisher Michael, Dennis Louise, Webster Matt, ‘Verifying Autonomous Systems’ Communications of the ACM, Vol. 56 No. 9, Pages 84-93, 10.1145/2494558

Fisher, 1991, Disambiguating anthropomorphism: An interdisciplinary review. J.A. Fisher Perspectives in Ethology, 9 (1991), pp. 49-85.

Fleming S M, Weil R S, Nagy Z, Dolan R J, and Rees G, ‘Relating introspective accuracy to individual differences in brain structure’ [2012] Science 329, 1541–1544. doi: 10.1126/science.1191883.

Franklin S and Graesser A, ‘Is it an agent or just a program? a taxonomy for autonomous agents’ [1997] Müller J.P., Wooldridge M.J., Jennings N.R. (eds) Intelligent Agents III Agent Theories, Architectures, and Languages. ATAL 1996. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 1193. Springer, Berlin, Heidelberg.

Franklin S and Graesser A, ‘Is it an agent or just a program? a taxonomy for autonomous agents’ [1997] Müller J.P., Wooldridge M.J., Jennings N.R. (eds) Intelligent Agents III Agent Theories, Architectures, and Languages. ATAL 1996. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 1193. Springer, Berlin, Heidelberg

Frost-Nielsen Per Marius, 'Bringing Military Conduct out of the Shadow of Law: Towards a Holistic Understanding of Rules of Engagement' [2018], Journal of Military Ethics, 17:1, 21-35, DOI: 10.1080/15027570.2018.1503217

Frowe Helen, the Ethics of war and Peace (2nd Edn, Routledge, 2016)

Fuchs Stephan, 'Beyond Agency', [2001] University of Virginia

Gabriel Iason, 'Artificial Intelligence, Values, and Alignment, Minds and Machines' (2020) 30:411–437 < https://doi.org/10.1007/s11023-020-09539-2 > accessed 4 September 2024.

Gabriel, I, 'Artificial Intelligence, Values, and Alignment' (2020) Minds & Machines 30, 411–437 < https://doi.org/10.1007/s11023-020-09539-2 > accessed 4 September 2014.

Gallie W B, Philosophers of Peace and War (1978) Cambridge: Cambridge University Press

Galliott Jai, MacIntosh Duncan, and Ohlin Jens David, Lethal Autonomous Weapons. Re-Examining the Law and Ethics of Robotic Welfare (2021 Oxford University Press)

Gamez David, Human and Machine Consciousness (Cambridge UK, Open Book Publishers 2018)

Gardner H, (1999). Intelligence Reframed: Multiple Intelligences for the 21st Century. New York, NY: Basic Book

Gehring W J, B Goss B, Coles M G H, Meyer D E, and Donchin E, 'A neural system for error detection and compensation' [1993] Psychol. Sci. 385–390. doi: 10.1111/j.1467-9280.1993.tb00586.x.

Geismann G, '"World Peace''. Rational Idea and Reality. On the Principles of Kant's Political Philosophy', [1996] H. Oberer, ed., Kant: Analysen, Probleme, Kritik . Germany: Konigshausen und Neumann (286)

Gennaro R J, (2019). Consciousness, The Internet Encyclopedia of Philosophy, ISSN 2161-0002. Available online at: https://www.iep.utm.edu/consciou/

Gennaro Rocco J, ''The Neuroscience of Psychiatric Disorders and the Metaphysics of Consciousness' [2019] Pascual Ángel Gargiulo & Humberto Luis Mesones Arroyo (eds.), Psychiatry and Neuroscience Update: From Translational Research to a Humanistic Approach, Volume III. Cham, Switzerland: Springer

Gibbs Samuel, 'Elon Musk leads 116 experts calling for outright ban of killer robots' (The Guardian, Sun 20 Aug 2017) < https://www.theguardian.com/technology/2017/aug/20/elon-musk-killer-robots-experts-outright-ban-lethal-autonomous-weapons-war > accessed 20 August 2017.

Gibbs Samuel, ‘Elon Musk leads 116 experts calling for outright ban of killer robots’ (The Guardian, Sun 20 Aug 2017) < https://www.theguardian.com/technology/2017/aug/20/elon-musk-killer-robots-experts-outright-ban-lethal-autonomous-weapons-war > accessed 20 August 2017

Gilovich T, Griffin D, and Kahneman D, ‘Heuristics and Biases’ [2002]: The Psychology of Intuitive Judgment. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511808098

Gleick Peter, ‘Protecting the environment in times of war’ (DownToEarth, 21 September 2019) < https://www.downtoearth.org.in/blog/climate-change/protecting-the-environment-in-times-of-war-66854> accessed 3 March 2020

Google Research People, 'Jeffrey Dean' (Google Research) <https://research.google/people/jeff/ > accessed on 10 October 2020

Gosseries O, Di H, Laureys S,and Boly M, ‘Measuring consciousness in severely damaged brains’ [2014] Annu. Rev. Neurosci. 37, 457–478. doi: 10.1146/annurev-neuro-062012-170339.

Gov.uk, ‘Product safety for manufacturers’ (Gov.uk, 30 October 2015) <https://www.gov.uk/guidance/product-safety-for-manufacturers > accessed on 5 November 2016

Gov.UK, ‘UK Defence Standardization’ (Gov.UK, 30 January 2024) < https://www.gov.uk/guidance/uk-defence-standardization > accessed 4 September 2024.

Gregersen Erik, "History of Technology Timeline" (Encyclopedia Britannica, 15 Jan. 2019) < https://www.britannica.com/story/history-of-technology-timeline> accessed 5 March 2019

Gronlund Kirsten, ‘State of AI: Artificial Intelligence, the Military and Increasingly Autonomous Weapons’ (Future of Life.org, 9 May 2019) < https://futureoflife.org/resource/state-of-ai/?cn-reloaded=1 > accessed 8 December 2021.

Grossberg Stephen, ‘Adaptive Resonance Theory: How a brain learns to consciously attend, learn, and recognize a changing world’ (2013) Neural Networks, Volume 37 < https://doi.org/10.1016/j.neunet.2012.09.017 > accessed 4 September 2024.

Gunkel D J, ‘Robot Rights’ [2018] MIT Press. doi: 10.7551/mitpress/11444.001.0001.

Hacker Donald E, ‘The Application of Prisoner-Of-War Status to Guerrillas Under the First Protocol Additional to the Geneva Conventions of 1949’ (1978) Boston College International and Comparative Law Review Volume 2 Issue 1 Article 7 1-1-1978 < https://core.ac.uk/download/pdf/80399938.pdf > accessed 10 November 2017

Hadji-Janev M and Hristovski K, 'BEYOND THE FOG: AUTONOMOUS WEAPON SYSTEMS IN THE CONTEXT OF THE INTERNATIONAL LAW OF ARMED CONFLICTS' (2017) Jurimetrics Journal of Law, Science and Technology, 57(3), 325+, < https://link.gale.com/apps/doc/A517345660/ITOF?u=griffith&sid=bookmark-ITOF&xid=2288b33d > accessed 6 Sep 2019.

Hadji-Janev M, and Hristovski K, 'BEYOND THE FOG: AUTONOMOUS WEAPON SYSTEMS IN THE CONTEXT OF THE INTERNATIONAL LAW OF ARMED CONFLICTS' (2017) Jurimetrics Journal of Law, Science and Technology, 57(3), 325+, < https://link.gale.com/apps/doc/A517345660/ITOF?u=griffith&sid=bookmark-ITOF&xid=2288b33d > accessed 6 Sep 2019

Haenlein M and Kaplan A, ‘A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence’ (2019) California Management Review, 61(4), 5-14. < https://doi.org/10.1177/0008125619864925 > accessed 4 September 2024.

Haladjian H H, and Montemayor C, ‘Artificial consciousness and the consciousness-attention dissociation’ [2016] Conscious. Cogn. 45, 210–225. doi: 10.1016/j.concog.2016.08.011.

Hallevy G (2013) When robots kill: artificial intelligence under criminal law. Northeastern University Press, Boston

Hallevy Gabriel, ‘Virtual Criminal Responsibility’ (8 May 2011) <https://ssrn.com/abstract=1835362> accessed on 17 November 2017.

Hardesty Larry, ‘Making computers explain themselves’ (MIT News, 27 October 2016) <http://news.mit.edu/2016/making-computers-explain-themselves-machine-learning-1028 > accessed on 24 April 2017

Harris J, ‘Wonderwoman and Superman,’ (1992) Oxford.

Harris John, Wonderwoman and Superman (1992) Oxford.

Havens John, ‘The ethics of AI: how to stop your robot cooking your cat’ (The Guardian, 23 June 2015) <https://www.theguardian.com/sustainable-business/2015/jun/23/the-ethics-of-ai-how-to-stop-your-robot-cooking-your-cat > accessed on 6 November 2016

Hegel G W F, ‘Philosophy of Right’ (2001) Transition (Vol. 1), Kitchener, Batoche Books Limited.

Heisenberg M, ‘Is Free Will an Illusion?’ [2009] Nature, 459: 164-165.

Hendriks B, Meerbeek B, Boess S, Pauws S, and Sonneveld M, ‘Robot vacuum cleaner personality and behavior’ [2011] Int. J. Soc. Robots 3, 187–195. doi: 10.1007/s12369-010-0084-5.

Hernandez Jeraldine, ‘Sex Robots: Negative Impact Towards Society’ (2018) Augustana Digital Commons, Winter 4-16-2018 < https://digitalcommons.augustana.edu/cgi/viewcontent.cgi?article=1014&context=ethicscontest> accessed 10 March 2019

Heyns Christof, ‘United Nations Report of the Special Rapporteur on extrajudicial, summary or arbitrary executions’ (2013) A/HRC/23/47

HFT, 'Mental Capacity Act (MCA)' (HFT, 2020) <https://www.hft.org.uk/resources-and-guidance/disability-rights-and-legal/mental-capacity-act/?gclid=EAIaIQobChMI9_7_sp2_8QIVkb7tCh2cSAY1EAAYASAAEgKIafD_BwE > accessed 10 March 2020

Hildt Elisabeth, ‘Artificial Intelligence: Does Consciousness Matter?’ (2019) Frontiers in Psychology Vol 1 <URL=https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01535> accessed 3 March 2020

Holder Chris, Khurana Vikram, Harrison Faye, and Jacobs Louisa, ‘Robotics and law: Key legal and regulatory implications of the robotics age (Part I of II)’ [2016] Computer Law & Security Review, Volume 32, Issue 3.

Holder Chris, Vikram Khurana, Faye Harrison, and Louisa Jacobs, ‘Robotics and law: Key legal and regulatory implications of the robotics age (Part I of II)’ [2016] Computer Law & Security Review, Volume 32, Issue 3

Holmes Marcia, ‘Hiding in Plain Sight’ (BBK, 29 June 2015) < http://www7.bbk.ac.uk/hiddenpersuaders/blog/hiding-plain-sight/ > accessed 4 September 2024.

Hooker J, ‘Autonomous Machines Are the Best Kind, Because They Are Ethical’, (2016), Carnegie Mellon University < http://public.tepper.cmu.edu/jnh/agencyPost2.pdf > Accessed on 14 December 2017

Horowitz Michael C., 'When speed kills: Lethal autonomous weapon systems, deterrence and stability' [2019] Journal of Strategic Studies, 42:6, 764-788, DOI: 10.1080/01402390.2019.1621174

Horrigan-Kelly M, Millar M, & Dowling M, ‘Understanding the Key Tenets of Heidegger’s Philosophy for Interpretive Phenomenological Research’ (2016) International Journal of Qualitative Methods, 15(1) < https://doi.org/10.1177/1609406916680634 > accessed 4 September 2024.

House of Commons Defence Committee, ‘Remote Control: Remotely Piloted Air Systems - current and future UK use’ (2014) House of Commons Sixth Special Report of Session 2014–15 < https://publications.parliament.uk/pa/cm201415/cmselect/cmdfence/611/611.pdf > accessed 17 November 2013

House of Commons Defence Committee, ‘Remote Control: Remotely Piloted Air Systems - current and future UK use’ (2014) House of Commons Sixth Special Report of Session 2014–15 < https://publications.parliament.uk/pa/cm201415/cmselect/cmdfence/611/611.pdf > accessed 17 November 2013

House of Lords Constitution Committee, “Constitutional arrangements for the use of armed force” Second Report, 2013-4 Session <https://publications.parliament.uk/pa/ld201314/ldselect/ldconst/46/46.pdf> accessed 10 November 2017

House of Lords Constitution Committee, Second Report, 2013-4 Session, “Constitutional arrangements for the use of armed force”

House of Parliament, Defence Committee, ‘Written evidence from the Ministry of Defence’ (Parliament.co.uk, 2013) < ‘https://publications.parliament.uk/pa/cm201314/cmselect/cmdfence/772/772vw02.htm > accessed 10 November 2017.

Houses of Parliament Paper, ‘Autonomous Road Vehicles’ (2013) Number 443 < http://researchbriefings.files.parliament.uk/documents/POST-PN-443/POST-PN-443.pdf > accessed on 11 November 2016

Hu Ying, Robot Criminals, (2019) U. Mich. J.L. Reform, Volume 52, issue 2, p. 487-531, < https://prospectusmjlr.files.wordpress .com/2019/04/robot-criminals.pdf > accessed 10 March 2020

Human Rights Watch, ‘Advancing The Debate on Killer Robots: 12 Key Arguments for a Pre-emptive Ban on Fully Autonomous Weapons’ (HRW, May 2014) < https://www.hrw.org/sites/default/files/related_material/Advancing%20the%20Debate_8May2014_Final.pdf > accessed in July 2016

Human Rights Watch, ‘Killer Robots and the Concept of Meaningful Human Control’ (2016) <https://www.hrw.org/news/2016/04/11/killer-robots-and-concept-meaningful-human-control > accessed on 6 November 2016.

Human Rights Watch, ‘Women’s Rights in Iran’ (Human Rights Watch, 28 October 2015) < https://www.hrw.org/news/2015/10/28/womens-rights-iran > accessed 10 November 2017

Hussain Naveed, Haddad Faris, Lester Robin, ‘DATA SCIENCE & DEEP LEARNING ABC…’ (2018) eBook v1.0 < https://meetnavpk.github.io/resources/ebook/pdf/Data_Science_&_Deep_Learning_ABC.pdf > accessed 3 March 2020.

Hutchinson Terry and Duncan Nigel, ‘Defining and Describing What We Do: Doctrinal Legal Research’ [2012] Deakin Law Review 17. 83-119.

IBM Research, ‘What’s Next in AI is foundation models at scale’ (IBM, 2025) < https://research.ibm.com/artificial-intelligence > accessed 4 January 2025

IBM Website, 'What is machine learning?' (IBM.COM) <https://www.ibm.com/topics/machine-learning?mhsrc=ibmsearch_a&mhq=what%20is%20machine%20learning > accessed 19 October 2020

IBM, ‘Biggest Data Challenges We Might Not Even Know’ (IBM, 2018) < https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/ > accessed 9 May 2019

IBM, ‘IBM Watson to Watsonx’ (IBM, 2024) < https://www.ibm.com/watson?utm_content=SRCWW&p1=Search&p4=43700080376796564&p5=p&p9=58700008735428981&gad_source=1&gbraid=0AAAAAoS6_Rcu9_aAPTO1_7sLHdSUNLQ7U&gclid=CjwKCAjw68K4BhAuEiwAylp3krqhOUBDfHoNaTgaZ_QW1boUqB2rQndEWbVL392XtEKsVSwEUjzpdxoCfQYQAvD_BwE&gclsrc=aw.ds > accessed 4 September 2024.

IBM, ‘What is Deep Learning?’ (IBM.com) <https://www.ibm.com/topics/deep-learning > accessed 20 August 2017

ICRC Advisory Service, ‘What is International Humanitarian Law?’ (2004) ICRC < https://www.icrc.org/sites/default/files/document/file_list/what-is-ihl-factsheet.pdf > accessed 4 September 2024.

ICRC, 'Lesson 10, The Law of Armed Conflict' (ICRC, 2002) <https://www.icrc.org/en/doc/assets/files/other/law10_final.pdf> accessed 10 November 2017

ICRC, 'Article 51 - Enlistment. Labour' (ICRC.org, 2024) < https://ihl-databases.icrc.org/en/ihl-treaties/gciv-1949/article-51> accessed 4 September 2024.

ICRC, 'Autonomous Weapon Systems Technical, Military, Legal and Humanitarian Aspects' (ICRC, 2014) <https://reliefweb.int/sites/reliefweb.int/files/resources/4221-002-autonomous-weapons-systems-full-report%20%281%29.pdf > accessed 17 November 2017

ICRC, 'Convention (IV) relative to the Protection of Civilian Persons in Time of War. Geneva, 12 August 1949' (ICRC, 2023) <https://ihl-databases.icrc.org/ihl/385ec082b509e76c41256739003e636d/6756482d86146898c125641e004aa3c5 > accessed 10 March 2020

ICRC, 'Convention (IV) respecting the Laws and Customs of War on Land and its annex: Regulations concerning the Laws and Customs of War on Land. The Hague, 18 October 1907' (ICRC, International Peace Conference 1907) < https://ihl-databases.icrc.org/en/ihl-treaties/hague-conv-iv-1907 > accessed 10 November 2017

ICRC, 'International Humanitarian Law and the challenges of contemporary armed conflicts' (ICRC, 2015) < https://www.icrc.org/en/document/international-humanitarian-law-and-challenges-contemporary-armed-conflicts > accessed 17 November 2017

ICRC, 'Introduction' (ICRC, 2020) <https://ihl-databases.icrc.org/ihl/INTRO/120?OpenDocument > accessed 3 March 2022

ICRC, 'Martens Clause' (ICRC.org, 2024) < https://casebook.icrc.org/a_to_z/glossary/martens-clause > accessed 4 September 2024.

ICRC, 'Proportionality' (ICRC.org, 2024) < https://casebook.icrc.org/a_to_z/glossary/proportionality > accessed 4 September 2024.

ICRC, 'Surrender' (ICRC.org, 2024) < https://casebook.icrc.org/a_to_z/glossary/surrender#:~:text=In%20international%20law%2C%20an%20isolated,perfidy%20and%20is%20therefore%20forbidden > accessed 4 September 2024.

ICRC, 'Target Verification' (ICRC) < https://ihl-databases.icrc.org/pt/customary-ihl/v2/rule16 > accessed 6 June 2019

ICRC, ‘The Fundamental Principles of the International Red Cross and Red Crescent Movement’ (2015) 4046/002 08.2015 5000 < https://www.icrc.org/sites/default/files/topic/file_plus_list/4046-the_fundamental_principles_of_the_international_red_cross_and_red_crescent_movement.pdf > accessed 4 September 2024.

ICRC, ‘The Principles of Humanity and Necessity’ (2023) ICRC < https://www.icrc.org/sites/default/files/wysiwyg/war-and-law/02_humanity_and_necessity-0.pdf > accessed 4 September 2024.

ICRC, ‘What are jus ad bellum and jus in bello?’, (ICRC, 22 January 2015) <https://www.icrc.org/en/document/what-are-jus-ad-bellum-and-jus-bello-0 > accessed 2 October 2019.

ICRC. ‘Military Necessity’ (ICRC.org, 2024) < https://casebook.icrc.org/a_to_z/glossary/military necessity#:~:text=The%20“principle%20of%20military%20necessity,prohibited%20by%20international%20humanitarian%20law > accessed 4 September 2024.

International Humanitarian Law and the challenges of contemporary armed conflicts.

International Humanitarian Law Databases, 'Treaties and States Parties’ (ICRC, 2024) < https://ihl-databases.icrc.org/en/ihl-treaties/treaties-and-states-parties > accessed 4 September 2024.

International Journal of Criminology and Sociological Theory, ‘Towards a New Sociology of Genetics and Human Identity’ [2013] Vol. 6, No.3, June 2013, 68-80 68

iRobot, ‘A Roomba home is a cleaner home’ (iRobot, 2024) < https://www.irobot.co.uk/en_GB/roomba.html?_gl=1*xvxtlq*_up*MQ..*_ga*NDA1ODUwNDMzLjE3MjkxOTQ2ODc.*_ga_WNZ0ESVFE6*MTcyOTE5NDY4Ni4xLjAuMTcyOTE5NDY4Ni4wLjAuMA > accessed 4 September 2024.

Ishida Yoshiteru and Chiba Ryunosuke, 'Free Will and Turing Test with Multiple Agents: An Example of Chatbot Design' [2017] Procedia Computer Science, Volume 112, 2017, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2017.08.190 > accessed 3 March 2019.

Ishida Yoshiteru, Chiba Ryunosuke, 'Free Will and Turing Test with Multiple Agents: An Example of Chatbot Design' [2017] Procedia Computer Science, Volume 112, 2017, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2017.08.190 > accessed 3 March 2019

Israeli Defence Forces Website, ‘‘The IDF Sees Artificial Intelligence as the Key to Modern-Day Survival’ (IDF, 2017) <https://www.idf.il/en/minisites/technology-and-innovation/the-idf-sees-artificial-intelligence-as-the-key-to-modern-day-survival/#:~:text=The%20IDF's%20Sigma%20branch's%20purpose,the%20IDF%20up%20to%20date> accessed 17 November 2017

Jafar Alzubi, Anand Nayyar and Akshi Kumar, ‘Machine Learning from Theory to Algorithms: An Overview’ (2018) J. Phys.: Conf. Ser. 1142 012012 <https://iopscience.iop.org/article/10.1088/1742-6596/1142/1/012012/meta?gclid=CjwKCAjwsKqoBhBPEiwALrrqiP3vMzDx9JGz1TANRsXtG34CXZmnRZDEY75gv0B6nf6tXjitLH8tPBoCvS4QAvD_BwE> accessed 3 March 2020

Jai Galliott, MacIntosh Duncan, and Ohlin Jens David, Lethal Autonomous Weapons. Re-Examining the Law and Ethics of Robotic Welfare (2021 Oxford University Press)

Jane Wakefield, ‘MEPs vote on robots' legal status - and if a kill switch is required’ (BBC News Online, 12 January 2017) < http://www.bbc.co.uk/news/technology-38583360 > accessed on 12 January 2017

Janelia Research Campus, ‘FlyEM Project’ (Janelia Research, 2023) < https://www.janelia.org/project-team/flyem > accessed 20 June 2023

Jha Dr U C, Killer Robots: Lethal Autonomous Weapon Systems – Legal, Ethical and Moral Challenges, (VIJ Books (India) 2016)

Jha U C, Killer Robots: Lethal Autonomous Weapon Systems – Legal, Ethical and Moral Challenges, (VIJ Books (India) 2016).

Johnson Arthur T, ‘Consciousness for Artificial Intelligence?’ (IEEE Pulse, 19 March 2024) < https://www.embs.org/pulse/articles/consciousness-for-artificial-intelligence/ > accessed 4 September 2024.

Johnstone Megan-Jane, 'Bioethics: A Nursing Perspective' (7th Edition, Elsevier 2019)

Johnstone Megan-Jane, Bioethics: A Nursing Perspective (7th Edition, Elsevier 2019)

Jones Thomas M, ‘Ethical Decision Making by Individuals in Organizations: An Issue-Contingent Model’ (1991) The Academy of Management Review, vol. 16, no. 2, 1991, pp. 366–95. JSTOR < https://doi.org/10.2307/258867 > accessed 4 September 2024.

Juretzki Bjoern, ‘Stakeholder Workshop 23rd November Digitising European Industry – Pillar 4’ (2015) European Commission < https://ec.europa.eu/digital-single-market/en/news/digitising-european-industry-external-stakeholders-group-meeting > accessed on 24 November 2017

Kant I, Fundamental Principles of the Metaphysic of Morals, (1785) 1949th ed. New York, NY: L. A. Press, Ed.

Katz B, 'Why Saudi Arabia Giving A Robot Citizenship Is Firing People Up' (Smithsonian Magazine, 2 November 2017) < https://www.smithsonianmag.com/smart-news/saudi-arabia-gives-robot-citizenshipand-more-freedoms-human-women-180967007/ > accessed 2 November 2017.

Katz B, 'Why Saudi Arabia Giving A Robot Citizenship Is Firing People Up' (Smithsonian Magazine, 2 November 2017) < https://www.smithsonianmag.com/smart-news/saudi-arabia-gives-robot-citizenshipand-more-freedoms-human-women-180967007/ > accessed 2 November 2017

Kauffman L H and Varela F J, 'Form dynamics' [1980] J. Soc. Biol. Syst.3, 171–206. doi: 10.1016/0140-1750(80)90008-1.

Kauffman L H, 'Self-reference and recursive forms' [1987] J. Soc. Biol. Syst. 10, 53–72. doi: 10.1016/0140-1750(87)90034-0.

Kauffman L H, and Varela F J, 'Form dynamics' [1980] J. Soc. Biol. Syst.3, 171–206. doi: 10.1016/0140-1750(80)90008-1

Kiltonsway, 'Artificial Intelligence (AI) vs. Machine Learning vs. Deep Learning' (Kiltonsway, 30 June 2021) < https://kiltonsway.mystrikingly.com/blog/artificial-intelligence-ai-vs-machine-learning-vs-deep-learning >accessed 13 June 2022.

Kinouchi Y, and Mackin K, 'A basic architecture of an autonomous adaptive system with conscious-like function for a humanoid robot' [2018] Front. Robot. 5:30. doi: 10.3389/frobt.2018.00030.

Kitwood T, Dementia Reconsidered: The Person Comes First (Open University Press 1997).

Klein Alice, 'Tesla driver dies in first fatal autonomous car crash in US' (2016) < https://www.newscientist.com/article/2095740-tesla-driver-dies-in-first-fatal-autonomous-car-crash-in-us/ > accessed in July 2016

Koch Christof, 'Will Machines Ever Become Conscious?' (Scientific American, 1 December 2019) < https://www.scientificamerican.com/article/will-machines-ever-become-conscious/ > accessed 3 April 2020.

Lacher Andrew, Grabowski Robert and Cook Stephen, ‘Autonomy, Trust and Transportation’ (2014) www.aaai.org < https://www.aaai.org/ocs/index.php/SSS/SSS14/paper/view/7701 > accessed on 1 November 2017

Lackey D P, ‘Review of Pacifism and the Just War., by J. Teichman’ (1993). Noûs, 27(4), 546–548 < https://doi.org/10.2307/2215800 > accessed 3 March 2019

Lee Dave, ‘US opens investigation into Tesla after fatal crash’ (BBC News Online, 1 July 2016) < http://www.bbc.co.uk/news/technology-36680043 > accessed on 1 July 2016

Leenes R and Lucivero F, 'Laws on robots, laws by robots, laws in robots: regulating robot behaviour by design' [2014] Law Innov Technol 6(2).

Levine David K, ‘Economic and Game Theory: What is Game Theory?’ (UCLA, date unknown) < http://levine.sscnet.ucla.edu/general/whatis.htm > accessed on 11/11/2016

LexiNexis “Recklessness definition’ (LexisNexis, 2024) < https://www.lexisnexis.co.uk/legal/glossary/recklessness#:~:text=In%20essence%2C%20recklessness%20means%20the,Lords%20in%20R%20v%20G.> accessed 4 September 2024.

LexisNexis, ‘Criminal act or omission’ (LexisNexis, 2024) < https://www.lexisnexis.co.uk/legal/guidance/criminal-act-or-omission > accessed 19 March 2019.

Library of Congress, Federal Research Division, ‘About this Service’ (Federal Research Division, Library of Congress) <https://www.loc.gov/rr/frd/Military_Law/NT_major-war-criminals.html > accessed 6 June 2023

Liivoja Rain, McCormack Tim, Handbook of the Law of Armed Conflict (Routledge 2016)

Lijadi Anastasia Aldelina, ‘What are universally accepted human values that define ‘a good life’? Historical perspective of value theory’ (2019) WP-19-006 IIASA < https://pure.iiasa.ac.at/id/eprint/16049/1/WP-19-006.pdf > accessed 4 September 2024

Lim Hannah YeeFin, Autonomous Vehicles and the law (Edward Elgar Publishing 2018)

Lin Partick, Jenkins Ryan and Abney Keith, Robot Ethics 2.0 (2020, Oxford University Press) 38-47.

Lin Patrick, Keith Abney, and Ryan Jenkins, Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence (Oxford University Press 2017)

Lloyd Ian J Information Technology Law (4th Edn Oxford University Press 2004)

Luck Michael and d'Inverno Mark, 'A Formal Framework for Agency and Autonomy' (1995) aaai.org < www.aaai.org/Papers/ICMAS/1995/ICMAS95-034.pdf> accessed on 20 October 2017

Lycan W G, and Dennett, D. C. (1993). Consciousness Explained. Philos. Rev. 102:424. doi: 10.2307/2185913

Lynch Gerald, 'Forget the Turing test: driverless cars need to pass a driving test first' (2016) <http://www.techradar.com/news/forget-the-turing-test-driverless-cars-need-to-pass-a-driving-test-first > accessed 7 November 2016

Machina M, 'Risk, ambiguity, and the rank-dependence axioms' [2009] Am. Econ. Rev. 99, 385–392. doi: 10.1257/aer.99.1.385

Majeed A B A, 'Roboethics - Making Sense of Ethical Conundrums' (2017) Procedia Computer Science, Volume 105, 2017 < https://doi.org/10.1016/j.procs.2017.01.227 > accessed 19 March 2019

Manganotti P, Formaggio E, Del Felice A, Storti S F, Zamboni A, Bertoldo A, Fiaschi A, and Toffolo G M, 'Time-frequency analysis of short-lasting modulation of EEG induced by TMS during wake, sleep deprivation and sleep' (2013) Front Hum Neurosci. 2013 Nov 18;7:767 < https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3831717/ > accessed 9 March 2020.

Marr Bernard, 'What Is GPT-3 And Why Is It Revolutionizing Artificial Intelligence?' (Forbes, 5 October 2020) <https://www.forbes.com/sites/bernardmarr/2020/10/05/what-is-gpt-3-and-why-is-it-revolutionizing-artificial-intelligence/> accessed 5 October 2020

Mataric Maja J, The Robotics Primer, (The MIT Press Cambridge, Massachusetts London, England 2007)

Matter of Nonhuman Rights Project, Inc v Stanley [2015] NY Slip Op 31419(U)

Matthew 7:12, New Testament.

Matthia C, Herbert Angermeyer Matschinger, 'The effect of violent attacks by schizophrenic persons on the attitude of the public towards the mentally ill' (1996) Social Science & Medicine, Volume 43, Issue 12 < https://doi.org/10.1016/S0277-9536(96)00065-2 > accessed 17 November 2017.

Matthia, Matschinger Herbert Angermeyer, 'The effect of violent attacks by schizophrenic persons on the attitude of the public towards the mentally ill' (1996) Social Science & Medicine, Volume 43, Issue 12 < https://doi.org/10.1016/S0277-9536(96)00065-2 > accessed 17 November 2017

McConnell Mark, To Be a Machine: Adventures Among Cyborgs, Utopians, Hackers, and the Futurists Solving the Modest Problem of Death (Granta Books, 2018)

McFadden Christopher, 'A Brief History of Military Robots Including Autonomous Systems' (Interesting Engineering, 06 Nov 2018) <https://interestingengineering.com/innovation/a-brief-history-of-military-robots-including-autonomous-systems> accessed 5 March 2019.

McGoogan Cara, ‘Google's self-driving car involved in serious crash after van jumps a red light’ (The Telegraph, 2016) < http://www.telegraph.co.uk/technology/2016/09/26/googles-self-driving-car-involved-in-serious-crash-after-van-jum/ > accessed on 26 September 2016

McLaughlin Eugene and Muncie John, The Sage Dictionary of Criminology (3rd edn Sage Publications, 2013)

McMahan Jeff, Killing in War (Clarendon Press, Oxford, 2009)

Microsoft Research, ‘Artificial Intelligence’ (Microsoft, 2025) < https://www.microsoft.com/en-us/research/research-area/artificial-intelligence/? > accessed 4 January 2025

Mills & Reeves, ‘Why we should get use to the idea that self-driving cars will sometimes crash’ (2015) Mills & Reeve < https://www.mills-reeve.com/files/Uploads/Documents/Autonomous-Vehicles-Article-Is-it-an-ethical-or-a-legal%20question.pdf > accessed on 24 November 2017.

Mills & Reeves, ‘Why we should get use to the idea that self-driving cars will sometimes crash’ (2015) Mills & Reeve < https://www.mills-reeve.com/files/Uploads/Documents/Autonomous-Vehicles-Article-Is-it-an-ethical-or-a-legal%20question.pdf > accessed on 24 November 2017

Mills and Reeve, ‘Legalising autonomous vehicles: Why it’s a drug related question’ (2016) Mills-Reeve.com < https://www.mills-reeve.com/legalising-autonomous-vehicles/ > accessed 2 November 2011.

Ministry of Defence, 'Joint Concept Note 1/18 Human-Machine Teaming' (Ministry of Defence, May 2018) <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/709359/20180517-concepts_uk_human_machine_teaming_jcn_1_18.pdf> accessed 5 March 2019

Ministry of Defence, 'UK Terminology Supplement to NATOTerm (JDP 0-01.1)' (Ministry of Defence, 1 September 2011) <https://www.gov.uk/government/publications/jdp-0-01-1-

united-kingdom-supplement-to-the-nato-terminology-database> accessed1 November 2016

Ministry of Defence, ‘A Soldier’s Guide to The Law of Armed Conflict’ (2005) Ministry of Defence <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/619906/2017-04714.pdf > accessed 2 October 2018.

Ministry of Defence, ‘Defence Drone Strategy. The UK’s Approach to Defence Uncrewed Systems’ (2024) < https://assets.publishing.service.gov.uk/media/65d724022197b201e57fa708/Defence_Drone_Strategy_-_the_UK_s_approach_to_Defence_Uncrewed_Systems.pdf > accessed 4 September 2024.

Ministry of Defence, ‘UK weapon reviews’ (2016) <https://www.gov.uk/government/publications/uk-weapon-reviews > accessed 10 November 2017.

Ministry of Defence, ‘UK Weapons Review’ (Ministry of Defence, 11 March 2016) < https://assets.publishing.service.gov.uk/media/5a80bf5f40f0b62305b8cec5/20160308-UK_weapon_reviews.pdf > accessed 19 March 2019.

Ministry of Defence, JSP 815. Element 5: Supervision, Contracting and Control Activities (JSP 815) Ministry of Defence < https://assets.publishing.service.gov.uk/media/66e18438dd4e6b59f0cb2500/JSP_815__Element_5_Supervision__contracting_and_control_activities_v1.2.pdf > accessed 4 September 2024.

Mitchell Tom, ‘Machine Learning 1st Edition’ (1997) McGraw-Hill Science < https://www.cin.ufpe.br/~cavmj/Machine%20-%20Learning%20-%20Tom%20Mitchell.pdf> accessed 10 November 2017

Moore D, ‘Measuring new types of question-order effects: Additive and subtractive’ [2002] Public Opin. Quart. 66, 80–91. doi: 10.1086/338631

Morgan Jonathan, ‘Military Negligence: Reforming Tort Liability after Smith v. Ministry of Defence Paper presented to the House of Commons Defence Select Committee’ [November 2013] Corpus Christi College, University of Cambridge < https://www.biicl.org/files/6759_military_negligence_paper-_jonathan_morgan.pdf > accessed 10 March 2020.

Morselli Alessandro, ‘The Mutual Interdependence between Human Action and Social Structure in the Evolution of the Capitalist Economy, Microeconomics and Macroeconomics’ [2014] Vol. 2 No. 1, 2014, pp. 6-11. doi: 10.5923/j.m2economics.20140201.02.

Mracek v. Bryn Mawr Hospital, March 11, 2009: <http://pa.findacase.com/research/wfrmDocViewer.aspx/xq/fac.20090311_0000339.EPA.htm/qx > accessed on 6 November 2016

Nakamura Kuninori, ‘The Footbridge Dilemma Reflects More Utilitarian Thinking Than The Trolley Dilemma: Effect Of Number Of Victims In Moral Dilemmas’ [undated] Tokyo Institute of Technology

National Cyber Security Centre, ‘Yahoo data breach: NCSC response’ (NCSC, 3 October 2017) <https://www.ncsc.gov.uk/news/yahoo-data-breach-ncsc-response > accessed 17 November 2019

Neural Link, ‘About’ (Neural Link, 2019) < https://neuralink.com/about/ > accessed 5 March 2020

NHS England, ‘NHS England business continuity management toolkit case study: WannaCry attack’ (NHS England, 21 April 2023) < https://www.england.nhs.uk/long-read/case-study-wannacry-attack/ > accessed 4 September 2024.

Niiniluoto Ilkka, 'Realism in Ontology' (2003) Critical Scientific Realism, Oxford Academic, 1 Nov. 2003 < https://doi.org/10.1093/0199251614.003.0002 > accessed 4 September 2024.

Nilsson N J, ‘The Quest for Artificial Intelligence’ (2009) Cambridge University Press < https://ai.stanford.edu/~nilsson/QAI/qai.pdf > accessed 10 November 2017

Nisbet Robert, Miner Gary, and Yale Ken, ‘Chapter 7 - Basic Algorithms for Data Mining: A Brief Overview’ (2018) Handbook of Statistical Analysis and Data Mining Applications (Second Edition), < https://doi.org/10.1016/B978-0-12-416632-5.00007-4 > accessed 4 September 2024.

Norbert Wiener (1894 – 1964) was an American mathematician and philosopher and credited as being one of the first to theorize about machine intelligence.

OpenAI, <https://openai.com/ > accessed 18 July 2022

Orend Brian (2004) Kant's ethics of war and peace, Journal of Military Ethics, 3:2, 161-177.

Orend Brian, War and International Justice: A Kantian Perspective (Wilfrid Laurier Univ. Press, 30 Oct 2010)

Owen A M, Coleman M R, Boly M, Davis M H, Laureys S, and Pickard J D, (2006) 'Detecting awareness in the vegetative state' [2006] Science 313:1402. doi: 10.1126/science.1130197.

Owen Tim and Owen Julie, 'Virtual Criminology: Insights from genetic-social science and Heidegger' [2015] Journal of Theoretical and Philosophical Criminology 7.17-31.

Owen Tim,

'Cyber-Violence: Towards a Predictive Model, Drawing upon Genetics, Psychology and Neur oscience' (2016) < clok.uclan.ac.uk/16131/1/40256-50104-1-PB-1.pdf> accessed on 6 November 2016

Owen Tim, and Jessica Marshall, Rethinking Cybercrime. Critical Debates (Palgrave Macmillan, 2021)

Owen Tim, Crime, Genes, Neuroscience and Cyberspace (Palgrave Macmillan 2017)

P Manganotti, E Formaggio, A Del Felice, SF Storti, A Zamboni, A Bertoldo, A Fiaschi, and GM Toffolo, 'Time-frequency analysis of short-lasting modulation of EEG induced by TMS during wake, sleep deprivation and sleep' (2013) Front Hum Neurosci. 2013 Nov 18;7:767 < https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3831717/ > accessed 9 March 2020

Pagallo Ugo, The Laws of Robots: Crimes, Contracts, and Torts (2013 edn, Springer.

Parliament, '5 Command Issues: Mission Command' (2004) Parliament.uk <https://publications.parliament.uk/pa/cm200304/cmselect/cmdfence/465/46508.htm> accessed 20 February 2017.

Parliament, 'Human rights law and International Humanitarian Law' (2014) Parliament.uk <https://publications.parliament.uk/pa/cm201314/cmselect/cmdfence/931/93106.htm> accessed 20 February 2017

Parliamentary Business, 'UK Armed Forces Personnel and the Legal Framework for Future Operations: Human rights law and International Humanitarian Law' (Parliament.UK, 2 April 2014) < https://publications.parliament.uk/pa/cm201314/cmselect/cmdfence/931/93106.htm >accessed 10 November 2017

Pictet Jean, 'The New Geneva Conventions for the Protection of War Victims' [1951] The American Journal of International Law, 45 (3): 462–475,

Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977.

QinetiQ, ‘MAARS weaponized robot’ (QinetiQ, 2023) <https://www.qinetiq.com/en/capabilities/robotics-and-autonomy/maars-weaponized-robot > accessed 20 June 2023.

Rayapati Naga, 'Artificial Intelligence Beyond Deep Neural Networks' (Forbes, 26 March 2019) <https://www.forbes.com/sites/forbestechcouncil/2019/03/26/artificial-intelligence-beyond-deep-neural-networks> accessed 5 March 2019

Renic Neil C, Asymmetric Killing (Oxford University Press 2020)

Restorative Justice Council, ‘What is restorative justice?’ (Restorative Justice Council, 2016) < https://restorativejustice.org.uk/what-restorative-justice > accessed 10 March 2020

Reynolds Emily ‘The agony of Sophia, the world's first robot citizen condemned to a lifeless career in marketing’ (Wired, Science, 01.06.2018) < https://www.wired.co.uk/article/sophia-robot-citizen-womens-rights-detriot-become-human-hanson-robotics > accessed 10 March 2019

Richards Neil M and Smart William D, ‘How Should the Law Think About Robots?’ (10 May 2013) < https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2263363> accessed on 17th November 2017.

Richards, Neil M. and Smart, William D, ‘How Should the Law Think About Robots?’ (10 May 2013). < https://ssrn.com/abstract=2263363>

Robotics Open Letter, ‘Open Letter to the European Commission Artificial Intelligence and Robotics’ <http://www.robotics-openletter.eu/ > accessed 10 November 2017

Roland C Mracek v. Bryn Mawr Hospital; Intuitive Surgical, Inc., United States Court of Appeals, Third Circuit. Submitted Under Third Circuit LAR 34.1(a) January 15, 2010. Opinion Filed: January 28, 2010

Rosanova M, Gosseries O, Casarotto S, Boly M, Casali A G, Bruno M A, et al. ‘Recovery of cortical effective connectivity and recovery of consciousness in vegetative patients’ [2012] Brain 135, 1308–1320. doi: 10.1093/brain/awr340.

Russell Bertrand, “The Ethics of War” (1915) International Journal of Ethics, vol. 25, no. 2, 1915, pp. 127–42 < http://www.jstor.org/stable/2376578 > accessed 24 May 2019

Sales P, ‘CONSTITUTIONAL VALUES IN THE COMMON LAW OF OBLIGATIONS’ (2024) The Cambridge Law Journal, 83(1) < https://www.cambridge.org/core/journals/cambridge-law-journal/article/constitutional-values-in-the-common-law-of-obligations/10695D32CEDAA3E2EEC391C212AF3925 > accessed 4 September 2024.

Samuels Thomas, Product Liability (2017 Westlaw)

Sandberg Anders, ‘Law-abiding robots?’, (2016) University of Oxford < https://www.oxfordmartin.ox.ac.uk/opinion/view/340 > accessed on 24 October 2017.

Saracco Roberto, Mason Dambrot Raj Madhavan, S., Derrick de Kerchove, and Tom Coughlin, ‘Symbiotic Autonomous Systems An FDC Initiative, White Paper’ (2017) IEEE.org <https://digitalreality.ieee.org/images/files/pdf/sas-white-paper-final-nov12-2017.pdf > accessed 13 April 2019

Saracco Roberto, ‘The evolution of … Machines’ (IEEE.org,October 27, 2021) https://cmte.ieee.org/futuredirections/2017/10/27/the-evolution-of-machines/ accessed 5 November 2021

Saracco Roberto, Madhavan Raj, Mason Dambrot S, de Kerchove Derrick, and Coughlin Tom, ‘Symbiotic Autonomous Systems An FDC Initiative, White Paper’ (2017) IEEE.org <https://digitalreality.ieee.org/images/files/pdf/sas-white-paper-final-nov12-2017.pdf > accessed 13 April 2019

Sarasso S, Rosanova M, Casali A G, Casarotto S, Fecchio M, Boly M, et al, ‘Quantifying cortical EEG responses to TMS in (Un)consciousness’ [2014] Clinical E. E. G. Neurosci. 45, 40–49. doi: 10.1177/1550059413513723.

Sassóli Marco, ‘Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified’ (2014) International Law Studies, US Naval college, INT’L L. STUD. 308 (2014) < https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=1017&context=ils > accessed 10 November 2017.

Sassóli Marco, ‘Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified’ (2014) International Law Studies, US Naval college, INT’L L. STUD. 308 (2014) < https://digital-commons.usnwc.edu/cgi/viewcontent.cgi?article=1017&context=ils > accessed 10 November 2017

Savirimuthu Joseph, “Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence” Patrick Lin, Keith Abney and Ryan Jenkins (eds), International Journal of Law and Information Technology, Volume 26, Issue 4, Winter 2018, Pages 337–346, <https://doi.org/10.1093/ijlit/eay011> accessed 9 January 2023

Savirimuthu Joseph, Do Algorithms Dream of 'Data' Without Bodies? (January 25, 2017) < https://ssrn.com/abstract=2905885 > Accessed 9 January 2023.

Savirimuthu, Joseph, "Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence" Patrick Lin, Keith Abney and Ryan Jenkins (eds), International Journal of Law and Information Technology, Volume 26, Issue 4, Winter 2018, Pages 337–346, <https://doi.org/10.1093/ijlit/eay011> accessed 9 January 2023

Saxena1 Abhay, Chang An-Yuan, Kulathuramaiyer Narayanan and Bhatt Ashutosh, "The Conceptual Framework of Articulated Human Intelligence in Perspective of Industry 4.0" [2019] Indian Journal of Science and Technology, Vol 12(6)

Scharre Paul, Autonomous Weapons and the Future of War: Army of None (Norton, 2019)

Schellekens Maurice, 'Self-driving cars and the chilling effect of liability law', [2015] Computer Law & Security Review, Volume 31, Issue 4, August 2015, Pages 506–517.

Scheutz M, "The inherent dangers of unidirectional emotional bonds between humans and social robots," [2011] Robot Ethics, The Ethical and Social Implications of Robotics, eds P. Lin, K. Abney, and G. A. Bekey (MIT Press)

Scholten Nina, 'The Robo-Criminal' [2019] Artificial Intelligence & Law (Fastcase) 263

SCMP, 'Tamagotchi grief counselling offered' (SCMP, 30 May 1997) < https://www.scmp.com/article/198008/tamagotchi-grief-counselling-offered > accessed 10 November 2017

Scruton Roger, 'Animal Rights and Wrongs' (2006) Continuum International Publishing Group Ltd

Searle J R, 'Minds, brains, and programs' [1980] Behav. Brain Sci. 3, 417–424. doi: 10.1017/S0140525X00005756.

Searle J R, 'Is the brain a digital computer?' [1990] Proc. Addres. Am. Philo. Assoc. 64, 21–37. doi: 10.2307/3130074

Searle J R, Consciousness and Language (Cambridge University Press 2002)

Sentencing Council, Child Cruelty Definitive Guideline (Definitive Guideline, Effective from 1 January 2019)

Sharkey Noel, 'Saying 'No!' to Lethal Autonomous Targeting' [2010] Journal of Military Ethics, 9:4, 369-383, DOI: 10.1080/15027570.2010.537903

Sharon Byrd B and Hruschka Joachim, 'Kant's Doctrine of Right: A Commentary' (Cambridge University Press; 1st edition,18 March 2010)

Shea N, and Frith C D, ‘Dual-process theories and consciousness: the case for “Type Zero” cognition’ [2016] Neurosci. Consci. 2016:niw005. doi: 10.1093/nc/niw005.

Sheila Brown, ‘The criminology of hybrids: Rethinking crime and law in technosocial networks’ (2006) Theoretical Criminology, 10(2), 223–244 < https://doi.org/10.1177/1362480606063140> accessed on 17 November 2017.

Sheridan T B, ‘Human-robot interaction: status and challenges’ [2016] Hum. Factors 58, 525–532. doi: 10.1177/0018720816644364.

Signorelli C M, ‘Types of cognition and its implications for future high-level cognitive machines’ (2017) AAAI Spring Symposium Series (Berkeley, CA) < http://aaai.org/ocs/index.php/SSS/SSS17/paper/view/15310 > accessed 3 March 2019.

Signorelli C M, and Arsiwalla X D, ‘Moral Dilemmas for Artificial Intelligence: a position paper on an application of Compositional Quantum Cognition’ [2018] Quantum Interaction. QI 2018. Lecture Notes in Computer Science (Nice).

Signorelli Camilo, ‘Can Computers Become Conscious and Overcome Humans?’ (2018) Hypothesis and Theory article Front. Robot. AI, Sec. Humanoid Robotics Volume 5 - 2018 < https://www.frontiersin.org/articles/10.3389/frobt.2018.00121/full > accessed 8 June 2019

Signorelli Camilo, ‘Can Computers Become Conscious and Overcome Humans?’ (2018) Hypothesis and Theory article Front. Robot. AI, Sec. Humanoid Robotics Volume 5 - 2018 < https://www.frontiersin.org/articles/10.3389/frobt.2018.00121/full > accessed 8 June 2019

Simester A P, G R Sullivan, J R Spencer QC, and Graham Virgo, Simester & Sullivan’s Criminal Law: Theory & Doctrine (4th ed, Hart Publishing 2010)

Simon Herbert, The Sciences of the Artificial (3rd edn, MIT Press 1996).

Singer Peter, Animal Liberation (First published 1975, Bodley Head 2015).

Singer Peter, Practical Ethics, 2nd Ed. Cambridge University Press (New York & Cambridge, U.K.: 1993).

Singer Peter, Practical Ethics, 2nd Ed. Cambridge University Press (New York & Cambridge, UK 1993)

Skyes Gresham and Matza David, ‘Techniques of Neutralization: A Theory of Delinquency’ American Sociological Review, Vol. 22, No. 6 (Dec. 1957), pp. 664-670

Smith Brad, and Browne Coral Ann, Tools and Weapons: The Promise and The Peril of the Digital Age (Hodder & Stoughton, 2019)

Smith J D 'The study of animal metacognition' [2009] Trends Cognit. Sci. 13, 389–396. doi: 10.1016/j.tics.2009.06.009

Snyder Kristy M, Ashitaka Yuki, Shimada Hiroyuki and Ulrich Jana E & Logan Gordon D, 'What skilled typists don't know about the QWERTY keyboard' (2013) Psychonomic Society, Inc. < https://link.springer.com/article/10.3758/s13414-013-0548-4 > accessed on 1 November 2017

SnyderKristy M & Ashitaka Yuki & Shimada Hiroyuki & Ulrich Jana E. & Logan Gordon D., 'What skilled typists don't know about the QWERTY keyboard' (2013) Psychonomic Society, Inc. < https://link.springer.com/article/10.3758/s13414-013-0548-4 > accessed on 1 November 2017

Social Care Institute for Excellence, 'Legislation relating to safeguarding adults' (December 2020) < https://www.scie.org.uk/key-social-care-legislation/safeguarding-adults > accessed 9 March 2020.

Solaiman S M, 'Legal personality of robots, corporations, idols and chimpanzees: a quest for legitimacy' (2017) Artif Intell Law 25, 155–179 (2017) <https://link.springer.com/content/pdf/10.1007/s10506-016-9192-3.pdf > accessed on 17th October 2018

Spinoza B, 'The Ethics' (1664/2009) Ethica Ordine Geometrica Demonstrata, New York, Dodo Press.

Stanford Encyclopedia of Philosophy, 'Kant's Social and Political Philosophy' (Stanford Encyclopedia of Philosophy, 2022) < https://plato.stanford.edu/entries/kant-social-political/#:~:text=The%20%E2%80%9Cuniversal%20principle%20of%20right,%E2%80%9D%20(6%3A230). > accessed 13 June 2022

Stanford Encyclopedia of Philosophy, 'The Chinese Room Argument' (Stanford Encyclopedia of Philosophy, First published Fri Mar 19, 2004; substantive revision Thu Feb 20, 2020) < https://plato.stanford.edu/entries/chinese-room/ > accessed 10 November 2017

Stanford Encyclopedia of Philosophy, 'Turing Test' (Stanford Encyclopedia of Philosophy, 4 October 2021) < https://plato.stanford.edu/entries/turing-test/ > accessed 4 September 2024.

Sternberg R J, (1997). The concept of intelligence and its role in lifelong learning and success. Am. Psychol. 52, 1030–1037. doi: 10.1037/0003-066X.52.10.1030

Stone P, Brooks R, Brynjolfsson E, Calo R, Etzioni O, Hager G, et al, ‘Artificial Intelligence and Life in 2030’ (2016) One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel. Stanford < http://ai100.stanford.edu/2016-report > accessed 10 November 2017.

Stone P, Brooks R, Brynjolfsson E, Calo R, Etzioni O, Hager G, et al, ‘Artificial Intelligence and Life in 2030’ (2016) One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel. Stanford < http://ai100.stanford.edu/2016-report > accessed 10 November 2017

Sweeney Joseph C, ‘The Just War Ethic in International Law’ (2003) 27 Fordham Int'l L.J. 1865 < https://ir.lawnet.fordham.edu/ilj/vol27/iss6/2 > accessed 4 September 2024

Synectics, ‘Evolution of Machine Learning’ (Synetices, 2018) < http://www.smdi.com/evolution-machine-learning > accessed 3 March 2020.

Taylor Rebecca, ‘Britain’s first self-driving cars will be unmarked to stop ‘bullying’ motorists trying to test their reactions by slamming on the brakes in front of them’ (Daily Mail Online, 30 November 2016) < http://www.dailymail.co.uk/news/article-3886972/Britain-s-self-driving-cars-unmarked-stop-bullying-motorists-trying-test-reactions-slamming-brakes-them.html#ixzz4ejDtNCuJ > accessed on 30 October 2016

Team Defence Information, ‘News Item’ (Team Defence Information, 2019) <https://www.teamdefence.info/news_item.php?sid=7d6b47c35c3c7120d6cca2244a28a793&newsitem=1000434> accessed 3 March 2018

Teichman Jenny, Pacifism And The Just War: A Philosophical Examination (Basil Blackwell 1986)

Tesla AI and Robotics, ‘AI and Robotics’ (Tesla, 2025) < https://www.tesla.com/en_gb/AI > accessed 4 January 2025

Teson F, ‘The Kantian Theory of International Law’ [1992] Columbia Law Review 92(1): 90

The Economist, ‘Morals and the machine’ (The Economist, 2 June 2012) <http://www.economist.com/node/21556234 > accessed on 6 November 2016

The Independent, ‘Stephen Hawking: 'Transcendence looks at the implications of artificial intelligence - but are we taking AI seriously enough?'’ (The Independent, 01 May 2014 < https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html > accessed 10 November 2017.

The Independent, ‘Stephen Hawking: 'Transcendence looks at the implications of artificial intelligence - but are we taking AI seriously enough?'’ (The Independent, 01 May 2014 < https://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html > accessed 10 November 2017

The Oxford Pocket Dictionary of Current English, ‘Machine’ (Encyclopedia.com, 26 May. 2023) <https://www.encyclopedia.com>. accessed 30 May 2023

The Rt Hon Stuart Andrew MP, ‘£2.5-million injection for drone swarms The Defence and Security Accelerator (DASA) has awarded £2.5-million to a consortium led by Blue Bear Systems Research Ltd to develop drone swarm technology’ (Ministry of Defence, Defence Science and Technology Laboratory, Defence and Security Accelerator, 28 March 2019) < https://www.gov.uk/government/news/25m-injection-for-drone-swarms> accessed 3 March 2022

Ticehurst Rupert, ‘The Martens Clause and the Laws of Armed Conflict’ (International Review of the Red Cross, April 1977) < https://www.icrc.org/eng/resources/documents/article/other/57jnhy.htm > accessed on 10 October 2016.

Tononi G and Koch C, ‘The neural correlates of consciousness: an update’ [2008] Ann. N. Y. Acad. Sci. 1124, 239–261. doi: 10.1196/annals.1440.004.

Tononi G, Boly M, Massimini M, and Koch C, ‘Integrated information theory: from consciousness to its physical substrate’ [2016] Nat. Rev. Neurosci. 17, 450–461. doi: 10.1038/nrn.2016.44

Tran Mark, ‘Girl starved to death while parents raised virtual child in online game’ (The Guardian, 5 Mar 2010) < https://www.theguardian.com/world/2010/mar/05/korean-girl-starved-online-game > accessed 10 November 2017

Traynor B J, and Singleton AB, ‘Nature versus nurture: death of a dogma, and the road ahead’ [2010] Neuron 68.

Tsagourias N, and Morrison, *International Humanitarian Law* (Cambridge University Press, 2018).

Turner Jacob, ‘Robot Rules. Regulating Artificial Intelligence’, 2019, Palgrave Macmillan.

Tzu Sun, The Art of War (Lionel Giles, 1910)

U.S Mission Geneva, ‘U.S Statement on Possible Options for Addressing the Humanitarian and International Security Challenges Posed by Emerging Technologies’ (U.S Mission to International Organisations in Geneva, 13 April 2018) <https://geneva.usmission.gov/2018/04/13/u-s-statement-on-possible-options-for-addressing-the-humanitarian-and-international-security-challenges-posed-by-emerging-technologies-in-the-area-of/ > accessed 3 March 2020

UK Autodrive <http://www.ukautodrive.com/ > accessed on 6 November 2016

UK Parliament, ‘The Law Governing Armed Conflict’ (Parliamentary Business, 22 July 2019), < https://publications.parliament.uk/pa/cm201719/cmselect/cmdfence/1224/122405.htm > accessed 4 September 2024.

UK Research and Innovation, ‘What is social science?’ (UKNI, 31 March 2022) https://www.ukri.org/who-we-are/esrc/what-is-social-science/qualitative-research/ > accessed 8 August 2024

Umbrello Steven, ‘No Machine Should Choose: Defending Human Dignity in the Age of Autonomous Weapons’ (Worldonfire.org, 19 September 2024) < https://www.wordonfire.org/articles/no-machine-should-choose-defending-human-dignity-in-the-age-of-autonomous-weapons/ > accessed 23 January 2025.

UN Ethics Office, ‘WHAT IS THE UN ETHICS OFFICE?’ (UN Ethics Office, 2024) < https://www.un.org/en/ethics/#:~:text=The%20UN%20Ethics%20Office%20promotes,and%20respect%20for%20human%20rights > accessed 4 September 2024.

United Nations, ‘1925 Geneva Protocol: Protocol for the Prohibition of the Use in War of Asphyxiating, Poisonous or Other Gases, and of Bacteriological Methods of Warfare’ (United Nations) < https://disarmament.unoda.org/wmd/bio/1925-geneva-protocol/ > accessed 3 March 2020

United Nations, ‘Geneva Convention for the Amelioration of the Condition of Wounded, Sick and Shipwrecked Member of Armed Forces at Sea of q12 August 1949’ (United Nations, 1949) < https://www.un.org/en/genocideprevention/documents/atrocity-crimes/Doc.31_GC-II-EN.pdf> accessed 10 November 2017

United Nations, ‘International Criminal Tribunal for the former Yugoslavia 1993’ (United Nations, 2017) < https://www.icty.org/ > accessed 20 March 2020

United Nations, ‘The Genocide Convention. 9 December 1948’ (United Nations, 2023) < https://www.un.org/en/genocideprevention/genocide-convention.shtml > accessed 10 March 2020

United Nations, ‘The Role of the United Nations in Addressing Emerging Technologies in the Area of Lethal Autonomous Weapons Systems’ (2018) Nos. 3 & 4 Vol. LV, "New Technologies: Where To?" < https://www.un.org/en/un-chronicle/role-united-nations-addressing-emerging-technologies-area-lethal-autonomous-weapons > accessed 3 March 2020

United Nations, ‘UN Diplomatic Conference Concludes in Rome with Decision to Establish Permanent International Criminal Court’ (UN Press Release, Press Release L/2889, 20 July 1998) <https://www.un.org/press/en/1998/19980720.l2889.html > accessed 3 March 2020

United Nations, ‘Universal values - peace, freedom, social progress, equal rights, human dignity - acutely needed, Secretary-General says at Tübingen University, Germany’ (2003) United Nations Press Release < https://www.un.org/press/en/2003/sgsm9076.doc.htm > accessed 3 September 2024.

United Nations, Office for Disarmament Affairs, ‘The Convention on Certain Conventional Weapons’ (United Nations, 2017) < https://www.unog.ch/80256EE600585943/(httpPages)/4F0DEF093B4860B4C1257180004B1B30?OpenDocument > accessed on 17 November 2017

United Nations, Office for Disarmament Affairs, ‘The Convention on Certain Conventional Weapons’ (United Nations, 2017) < https://www.unog.ch/80256EE600585943/(httpPages)/4F0DEF093B4860B4C1257180004B1B30?OpenDocument > accessed on 17 November 2017.

United Stated of America Before the Federal Trade Commission, In the Matter of Cambridge Analytica, LLC, a corporation, 182 3107 < https://www.ftc.gov/system/files/documents/cases/182_3107_cambridge_analytica_administrative_complaint_7-24-19.pdf > accessed 18 March 2020

United States Congressional Serial Set, ‘Issue 3794’ (United States Congressional) < United States Congressional serial set, Issue 3794, p. 785 > accessed 10 March 2022

Urquiza Esmeralda & Kotrschal Kurt, ‘The mind behind anthropomorphic thinking: Attribution of mental states to other species’ [2015] Animal Behaviour. 109. 167-176. 10.1016/j.anbehav.2015.08.011

US Department of Defense Directive NUMBER 3000.09 November 21, 2012 < https://fas.org/irp/doddir/dod/d3000_09.pdf > accessed on 24 October 2017.

US Department of Defense, Conduct of the Persian Gulf War, Final Report to Congress (1992) (Department of Defense Report) 613.

Van Gulick F, (2018). "Consciousness," in The Stanford Encyclopedia of Philosophy, ed E. N. Zalta. Available online at: https://plato.stanford.edu/entries/consciousness/

van Gulick Robert, Consciousness (Stanford Encyclopedia of Philosophy 2004)

Varela F J, 'A calculus for self-reference' [1975] Int. J. Gen. Syst. 2, 5–24.

Varela F J, and Goguen J A, 'The arithmetic of closure' [1978] Cybernet. 8, 291–324. doi: 10.1080/01969727808927587.

Varelius J 'The value of autonomy in medical ethics' [2006] Med Health Care Philos. 2006;9(3):377-88. doi: 10.1007/s11019-006-9000-z. Epub 2006 Oct 11. PMID: 17033883; PMCID: PMC2780686.

Varkey B 'Principles of Clinical Ethics and Their Application to Practice' (2020) Med Princ Pract. 2021;30(1):17-28 < .https://pmc.ncbi.nlm.nih.gov/articles/PMC7923912/#:~:text=Beneficence%2C%20nonmaleficence%2C%20autonomy%2C%20and%20justice%20constitute%20the%204%20principles,the%20latter%202%20evolved%20later > accessed 4 September 2024.

Velasquez Manuel, Andre Claire, Shanks Thomas, and Meyer Michael J, 'What is Ethics?' (Markkula Center for Applied Ethics, Santa Clara University, 1 January 2010) < https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/what-is-ethics/#:~:text=First%2C%20ethics%20refers%20to%20well,%2C%20fairness%2C%20or%20specific%20virtues > accessed 4 September 2024.

Vijayenthiran Viknesh, 'Mercedes is backtracking on claims its self-driving cars will kill pedestrians over passengers in close calls' (Business Insider UK, October 2016) < http://www.businessinsider.com/mercedes-denies-claim-its-driverless-car-will-prioritize-driver-safety-2016-10?IR=T > accessed on 1 November 2016

Vișan Cosmin, 'The Self-Referential Aspect of Consciousness' (2017) Journal of Consciousness Exploration & Research Vol 8 Iss 11 < https://philarchive.org/archive/VISTSA-2 > accessed 4 September 2024.

Walczak, Steven, Cerpa, Narciso, ‘Artificial Neural Networks’ (2003) Encyclopedia of Physical Science and Technology, 3rd Edn, Pages 631-645, <https://doi.org/10.1016/B0-12-227410-5/00837-1> accessed 20 August 2017

Wall Matthew, ‘Is this the year 'weaponised' AI bots do battle?’ (BBC Website, 5 January 2018) < www.bbc.co.uk/news/business-42559967 > accessed 5 January 2018

Wallach Wendell and Allen Colin, Moral Machines: Teaching Robots Right from Wrong: Teaching Robots Right from Wrong (2008) Oxford University Press

Wang Y, ‘The cognitive mechanisms and formal models of consciousness’ [2012] Int. J. Cognit. Inform. Nat. Intel. 6, 23–40. doi: 10.4018/jcini.2012040102.

Warren Mary, ‘On the Moral and Legal Status of Abortion’ The Monist, 57 (1973).

Weapons Law, ‘Glossary: martens clause’ (Weapons Law, date unknown) < http://www.weaponslaw.org/glossary/martens-clause > accessed on 1 November 2016

Wearver John Frank, Robots Are People Too (Praeger, 2014)

Weblearn,Oxford University, ‘Foundational Myths in the Law of war’ (Weblearn, Oxford University, 2020) < https://weblearn.ox.ac.uk/access/content/user/1044/MJIL%20July%202019%20-%20AR%20on%20Foundational%20Myths%20in%20the%20Laws%20of%20War.pdf > accessed 10 March 2020

Weller Chris, ‘The world's first artificially intelligent lawyer was just hired at a law firm’ (Business Insider UK, May 2016) < http://uk.businessinsider.com/the-worlds-first-artificially-intelligent-lawyer-gets-hired-2016-5?r=US&IR=T > accessed on 17 October 2016

Weng Y H, Chen C H and Sun CT, ‘Toward the human–robot co-existence society: on safety intelligence for next generation robots’ [2009] Int J Social Robot 1:267–282.

West Yorkshire Police, ‘Police Dog Bites’ (West Yorkshire Police, 2020) < https://www.westyorkshire.police.uk/sites/default/files/2020-06/police_dog_bites.pdf > accessed 20 June 2020.

Wiesel Elie, ‘Night,’ [1972] Penguin Books.

Williams Garrath, ‘Kant’s Account of Reason’, The Stanford Encyclopedia of Philosophy (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), <https://plato.stanford.edu/archives/fall2023/entries/kant-reason/> accessed 10 September 2023.

Williams H, 'Judgements on War: A Response' (1995) H. Robinson, ed., Proceedings of the Eighth International Kant Congress, Vol. 1, Part 3. Milwaukee, WN: Marquette University Press < https://www.pdcnet.org/kant1995/content/kant1995_1995_0001_0003_1385_1393 > accessed 3 March 2020

Wooldridge Micheal, The Road to Conscious Machines: The Story of AI (Pelican Books, 2020).

World Economic Forum, 'Emerging Technologies: Top 9 ethical issues in artificial intelligence' (World Economic Forum, 21 October 2016) <https://www.weforum.org/agenda/2016/10/top-10-ethical-issues-in-artificial-intelligence/> accessed 10 November 2017

Wu Stephen, 'Summary of Selected Robotics Liability Cases' (2010) Cooke Kobrick & Wu LLP < http://ftp.documation.com/references/ABA10a/PDfs/2_5.pdf > accessed on 24 November 2017

Young G, 'Personhood across disciplines: Applications to ethical theory and mental health ethics in Ethics, Medicine and Public Health' (2019) Ethics, Medicine and Public Health Volume 10, July–September < https://www.sciencedirect.com/topics/medicine-and-dentistry/personhood#:~:text=Personhood%20is%20a%20normative%20category,the%20person%20(after%20Kant) > accessed 4 September 2024.

Legislation

Balding v Lew-Ways Limited (1995) 159 JP 541

Brown and Son Ltd v Craiks Ltd [1970] 1 WLR 752

Coke v. Bumble – comments on some aspects of unlawful killing and its disposal Herschel

Cunningham [1957] 2 QB 396

Dalloway (1847) 2 Cox CC 273

Davis v Komatsu America Industries Corp., 42 S.W.3d 34 (Tenn. 2001) – US case

Donoghue v Stevenson [1932] AC 562

Evans [2009] EWCA Crim 650; [2009] 1 WLR 1999

G and another [2003] UKHL 50; [2003] 3 WLR 1060

Gibbins and Proctor (1918) 13 Cr App R 134

Hughes [2013] UKSC 56; [2013] 1 WLR

Judgement Smith and others (FC) (Appellants) v The Ministry of Defence (Respondent) Ellis (FC) (Respondent) v The Ministry of Defence (Appellant) Allbutt and others (FC) (Respondents) v The Ministry of Defence (Appellant), [2013] UKSC 41 <https://www.supremecourt.uk/cases/docs/uksc-2012-0249-judgment.pdf> accessed 3 March 2020

Lambert v Lewis [1982] AC 225

Matter of Nonhuman Rights Project, Inc v Stanley [2015] NY Slip Op 31419(U)

Metropolitan Police Commissioner v Caldwell [1982] AC 341

Millar [1983] 2 AC 161

Nicholls (1874) 13 Cox CC 75

Pittwood (1902) 19 TLR 37

Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977

R v Dudley and Stephens (1884) 14 QBD 273

R v G [2004] 1 A.C. 1034

R v Jogee [2016] UKSC 8

R v Sergeant Alexander Wayne Blackman ("Marine A"), Case Reference: 2012CM00442 < https://www.judiciary.uk/wp-content/uploads/JCO/Documents/Judgments/r-v-blackman-marine-a-sentencing+remarks.pdf > accessed 3 March 2020

Smith and others (Appellants) v The Ministry of Defence (Respondent) Ellis (Respondent) v The Ministry of Defence (Appellant) Allbutt and others (Respondents) v The Ministry of Defence (Appellant) [2013] UKSC 41

Stephenson [1979] QB 695

Vaughan v Menlove (1837) 3 Bing NC 467

White [1910] 2 KB 124

Woollin [1998] UKHL 28; [1998] 3 WLR 382

Cases

Animal Welfare (Service Animals) Act 2019

Animal Welfare Act 2006

Armed Forces Act 2006

Armed Forces Act 2011

Armed Forces Act 2016

Article 1 of the European Convention on Human Rights

Article 10 of the European Convention on Human Rights

Article 2 of the European Convention on Human Rights

Article 36 of Additional Protocol I to the Geneva Conventions

Consumer Protection Act 1987

Consumer Rights Act 2015

Corporate Manslaughter and Corporate Homicide Act 2007

Female Genital Mutilation Act 2003

Free speech and hate speech under EU Article 10 of the European Convention on Human Rights

General Product Safety Regulations 2005/1803

Geneva Convention (IV) on Civilians, 1949 - Article 51

International Human Rights Law

Mental Capacity Act 2005

Mental Health Act 1983

Misrepresentation Act 1967

Modern Slavery Act 2015

The Armed Forces Act 2006

The Children and Young Persons Act 1933

The Consumer Protection Act 1987 (CPA 1987) is central to product liability, with Parts II and IV covering criminal liability. Tortious liability common law claims are founded upon Donoghue v Stevenson [1932] AC 562.

The Criminal Damage Act 1971

The Hague Convention of 1907

The Mental Health Act 1983 and Mental Capacity Act 2005

The Modern Slavery Act 2015