Research paper series

ASSER INSTITUTE
Centre for International & European Law

July 2022

Are programmers in or 'out of' control? The individual criminal responsibility of programmers of autonomous weapons and self-driving cars

Dr. Marta Bo

08

Centre for International & European Law
www.asser.nl

Forthcoming in: Gless, S. & Whalen-Bridge, H. (eds.), *Human-Robot Interaction in Law and its Narratives: Legal Blame, Criminal Law, and Procedure*, Cambridge University Press (forthcoming 2022)

Cite as: ASSER research paper 2022-08

Author Contact Details: m.bo@asser.nl

Abstract

The increasing use of autonomous systems technology in cars and weapons could lead to a rise of harmful incidents on the roads and in the battlefield potentially amounting to crimes. Such a rise has led to questions as to who is criminally responsible for these crimes – be it the users or the programmers? This chapter seeks to clarify the role of programmers in crimes committed with autonomous systems by focusing on the use of autonomous vehicles and autonomous weapons. In assessing whether a programmer could be criminally responsible for crimes committed with autonomous technology, it is necessary to determine whether the programmer had control over this technology. Risks inherent in the use of these autonomous technologies may allow for a programmer to escape criminal liability but some risks may be foreseeable and thus considered under the programmer's control. The central question is whether programmers exercise causal control over a chain of events leading to the commission of a crime. This chapter contends that programmers' control begins at the initial stage of the autonomous system development process but continues in the use phase, extending to the behaviour and effects of autonomous systems technology. Based on criminal responsibility requirements and causation theories, this chapter develops a notion of meaningful human control (MHC) that may function to trace back responsibility to the programmers who could understand, foresee, and anticipate the risk of a crime being committed with autonomous systems technology.

Keywords

International criminal law, autonomous weapons, criminal law, self-driving cars, criminal responsibility, programmers, AI, causation, artificial intelligence

Marta Bo, 'Are Programmers in or 'out of' Control? The Individual Criminal Responsibility of Programmers of Autonomous Weapons and Self-Driving Cars: in S. Gless, H. Whalen-Bridge (eds.), *Human-Robot Interaction in Law and its Narratives: Legal Blame, Criminal Law, and Procedure* (CUP: forthcoming, 2022)

*Are Programmers in or 'out of' Control? The Individual Criminal Responsibility of Programmers of Autonomous Weapons and Self-Driving Cars**

1. Introduction

In March 2018, a Volvo XC90 vehicle that was being used to test Uber's emerging automated vehicle technology hit and killed a pedestrian crossing a road in Tempe, Arizona.[1] At the time of the incident the vehicle was in 'autonomous mode' and the vehicle's operator, Rafaela Vasquez, was allegedly streaming television onto their mobile device.[2] In November 2019, the National Transportation Safety Board found that many factors contributed to the fatal incident, including failings from both the vehicle's operator and programmer of the autonomous system: Uber.[3] Nevertheless, despite Vasquez later being charged with negligent manslaughter in relation to the incident,[4] criminal investigations into Uber were discontinued in March 2019.[5] This instance is particularly emblematic of the current tendency to consider responsibility for actions and decisions of autonomous vehicles (AVs) as primary lying with users of these systems, and not programmers.[6]

In the military realm, similar issues have arisen. For example, it is alleged that in 2020 an autonomous drone system – the *STM Kargu-2* – may have been used during active hostilities in Libya.[7] It is purported that such autonomous weapons (AWs) were programmed to attack

* Dr Marta Bo is a Researcher at the University of Amsterdam-T.M.C. Asser Institute (The Hague); Associate Senior Researcher at SIPRI and a Postdoctoral Researcher at The Graduate Institute (Geneva) where she is a member of the Swiss National Science Foundation (SNSF)-funded project 'Lethal Autonomous Weapon Systems and War Crimes: Who Is to Bear Criminal Responsibility for Commission?' (Project 10001C_176435) which funded this research. I am grateful for the helpful discussions within the DILEMA team at the Asser Institute, the comments and feedbacks received on an earlier draft during the author workshop within the SNSF project 'Human-Robot Interaction in Law and its Narratives: Legal Blame, Criminal Law, and Procedure'. I thank James Patrick Sexton for his invaluable research assistance and helping me improve earlier drafts of this book chapter. All errors remain mine.

[1] S. Levin and J.C. Wong, 'Self-Driving Uber kills Arizona woman in first fatal crash involving pedestrian', *The Guardian (online)*, 19 March 2018, available at: www.theguardian.com/technology/2018/mar/19/uber-self-driving-car-kills-woman-arizona-tempe.

[2] L. Binding, 'Arizona Uber driver was 'streaming The Voice' moments before fatal crash', *Sky News*, 22 June 2018, available at: news.sky.com/story/arizona-uber-driver-was-streaming-the-voice-moments-before-fatal-crash-11413233. In this chapter I will use interchangeably the terms 'driver', 'occupant', 'operator' and 'user'.

[3] National Transportation Safety Board, 'Highway Accident Report: Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian Tempe, Arizona March 18, 2018', 19 November 2019, available at: www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf.

[4] *The State of Arizona vs. Rafael Stuart Vasquez*, Indictment 785 GJ 251, Superior Court of the State of Arizona in and for the County of Maricopa, 27 August 2020, available at: www.maricopacountyattorney.org/DocumentCenter/View/1724/Rafael-Vasquez-GJ-Indictment.

[5] BBC News, 'Uber "not criminally liable" for self-driving death', 6 March 2019, available at: www.bbc.co.uk/news/technology47468391#:~:text=Uber%20will%20not%20face%20criminal%20charges%20for%20a,car%27s%20back-up%20driver%20could%20still%20face%20criminal%20charges.

[6] Manufacturers of AVs often include responsibility clauses in their contracts with end-users. However, practice may vary as 'Volvo has already made public its willingness to accept full liability, whereas Tesla has stated that it will accept liability only for design failure', K. Grieman, 'Hard Drive Crash: An Examination of Liability for Self-Driving Vehicles', 3 *JIPITEC* 294 (2018), 300, para. 29.

[7] United Nations Security Council, 'Letter dated 8 March 2021 from the Panel of Experts on Libya established pursuant to resolution 1973 (2011) addressed to the President of the Security Council', 8 March 2021, S/2021/229, paras. 63-64.

targets without requiring data connectivity between the operator and the use of force.[8] Although such AWs technologies have not yet been widely used by militaries, governments, civil society, and academics have for several years debated their legal position, especially highlighting the importance of retaining 'meaningful human control' (MHC) in decision making processes to prevent potential 'responsibility gaps'.[9] When debating MHC over AWs as well as responsibility issues, users or deployers are more often scrutinised than programmers,[10] the latter being considered too far removed from the effects of AWs. However, programmers' responsibility increasingly features in policy and legal discussions, leaving many interpretative questions open.[11]

To fill this gap in the current debates, this Chapter seeks to clarify the role of programmers in crimes committed *with* (and *not by*) AVs and AWs (AV- and AW-related crimes). The origin of this problem is AI systems' incapability of fulfilling the *mens rea* (mental element) and *actus reus* (conduct element including its causally connected consequences) generally required by criminal law.[12] Thus, the criminal responsibility of programmers will be considered in terms of direct responsibility for commission (i.e., perpetrators or co-perpetrators)[13] rather than vicarious or joint responsibility for crimes committed *by* AI (considered as perpetrators of crimes). Moreover, programmers could, for example, be held responsible on the basis of participatory modes of responsibility, such as aiding or assisting users in the perpetration of a crime. Despite their relevance to a discussion on the responsibility of programmers, participatory modes of responsibility under national and international criminal law would require a separate analysis since they are each characterised by different *actus reus* and *mens rea* standards. Finally, it must be acknowledged that the term 'programmer' used for the purpose of this chapter is a simplification. The development of AVs and AWs entails the involvement of numerous actors, internal and external to tech companies, such as developers, programmers, data labellers, component manufacturers, software developers, and manufacturers. This might entail difficulties in individualising responsibility and/or a distribution of criminal responsibility, which could be captured by participatory modes or responsibility.

This Chapter will examine the criminal responsibility of programmers through two examples: AVs and AWs. Granted, there are some fundamental differences between AVs and AWs: AWs are intended to kill and are inherently dangerous, while AVs are not. However, AW use may unintentionally result in unlawful harmful incidents and killing. Thus, a common

[8] *Ibid.*, para. 63.

[9] *See* F.S. de Sio and J. van den Hoven, 'Meaningful Human Control over Autonomous Systems: A Philosophical Account', 5 *Frontiers in Robotics and AI* 1 (2018); Human Rights Watch, 'Killer Robots and the Concept of Meaningful Human Control: Memorandum to Convention on Conventional Weapons (CCW) Delegates', 11 April 2016, ICRC, 'Artificial intelligence and machine learning in armed conflict: A human-centred approach' (2019).

[10] B. Boutin and T. Woodcock, 'Aspects of Realizing (Meaningful) Human Control: Legal Perspective', in R. Geiß and H. Lahmann (eds.), *Research Handbook on Warfare and Artificial Intelligence* (Elgar: forthcoming), 9.

[11] M. Bo, L. Bruun, V. Boulanin, *Human responsibility for the development and use of autonomous weapon systems: Ensuring state responsibility and individual criminal responsibility for violations of international humanitarian law involving AWS*, (forthcoming 2022).

[12] *See* T.C. King, N. Aggarwal, M. Taddeo, and L. Floridi, 'Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions', 26 *Science and Engineering Ethics* 89 (2020), 95; *see contra* the work of G. Hallevy, 'The Criminal Liability of Artificial Intelligence Entities: from Science Fiction to Legal Social Control', 4 *Akron Intellectual Property Journal* 171 (2010).

[13] Direct commission or principal responsibility under international criminal law also includes joint commission and co-perpetration, G. Werle and F. Jessberger, *Principles of Internlational Criminal Law* (OUP: 2020), paras. 623-659. Co-perpetratorship is also a form of principal responsibility in German criminal law and is founded on the concept of 'control over whether and how the offense is carried out', T. Weigend, 'Germany', in K.J. Heller and M.D. Dubber (eds.), *The Handbook of Comparative Criminal Law* (Stanford: 2011), 265 and 266. There is no similar 'co-perpetration' mode of liability in the United States (US).

feature is the unlawful and direct harm to life and physical health that could arise from AV and AW use. Moreover, AVs are means of transport and as such this implies the presence of people onboard, which will not necessarily be a feature of AWs. Moreover, for both AVs and AWs object recognition technology[14] is often the source of incidents resulting in harm to individuals. The focus here are crimes against persons under national criminal law (i.e., manslaughter and negligent homicide) stemming from the use of AVs, and crimes against persons under international criminal law resulting from the use of AWs (i.e., war crimes against civilians under international law, such as found in the Rome Statute of the International Criminal Court (ICC)[15] and in the First Additional Protocol to the Geneva Conventions).[16]

A core issue is whether programmers could fulfil the *actus reus*, including the requirement of causation, of these crimes. Given the temporal and spatial gap between programmers' conduct and the possible intervention of other causes, a core challenge in ascribing criminal responsibility lies in determining a causal link between programmers' conduct and AV- and AW-related crimes. To determine this, it is necessary to delve into the technological aspects of AVs and AWs and consider when and which of their associated risks can or cannot be in principle imputable to a programmer.[17] Adopting a preliminary categorisation of AV- and AW-related risks based on programmers' alleged control (or lack of) over the *behaviour* and/or *effects* of AVs and AWs, Sections 2 and 3 are concerned with the different risks and incidents entailed by the use of AVs and AWs. Section 4 will then turn to the elements of the AV- and AW- related crimes focusing on causation tests and touching upon *mens rea*. Drawing from this analysis, Section 5 will turn to a notion of '*meaningful control*' over AVs and AWs that incorporates requirements for the ascription of criminal responsibility, and, in particular, causation criteria to determine under which conditions programmers exercise *causal control* over the unlawful *behaviour* and/or *effects* of AVs and AWs

2. Risks Posed by AVs and Programmers' Control[18]

Without seeking to identify all possible causes of AV-related incidents, Section 2 begins by identifying several risks associated with AVs: algorithms, data, users, vehicular communication technology, hacking, and the behaviour of bystanders. Some of these, such as those linked to supervised and unsupervised learning algorithms, are also applicable to AWs.

In order to demarcate a programmer's criminal responsibility, it is crucial to determine whether they ultimately had *control* over the *behaviour* and *effects* (i.e. navigation and some possible consequences) of AVs. Thus, the following sub-sections make a preliminary distinction of risks on the basis of the programmers' alleged *control* over them. While a notion of (meaningful) *control* which encompasses the requirement of causality in criminal law will be developed in Section 5, it is important at this point to anticipate that a fundamental threshold

[14] *See* below Sections 2 and 3.

[15] Rome Statute of the International Criminal Court (adopted 17 July 1998, entered into force 1 July 2002) 2187 UNTS 3. Hereinafter 'Rome Statute'.

[16] Protocol Additional to the Geneva Conventions of 12 August 1949 and Relating to the Protection of Victims of International Armed Conflicts (signed 8 June 1977, entered into force 7 December 1978) 1125 UNTS 3. Hereinafter 'AP I'.

[17] Some theories of causation recognise that causation in law is a matter of imputation', i.e a matter of imputing a result to a criminal conduct, P.K. Ryu, 'Causation in Criminal Law' 106 *U. Pa. L. Rev.* 773 (1958), 785, 795 and 796.

[18] In the context of AVs, since many major car manufacturers have produced and programmed their own AVs, the responsibility of manufacturers and programmers might overlap, *see* Grieman, *supra* note 6, 300.

for establishing the required causal nexus between a conduct and harm is whether a programmer could understand, foresee and anticipate a certain risk and whether the risk that materialised was within the scope of the programmer's functional obligations.[19]

2.1 Are Programmers in Control of Algorithm and Data-Related Risks in AVs?

Before turning to the risks and failures that might lie in the phase of algorithm design and thus be potentially considered under programmers' *control*, this sub-section, first, describes the tasks being made increasingly autonomous in AVs and, second, some of the rules to be coded to this end.

The main task of AVs is navigation, which can be understood as the AV's *behaviour* as well as the algorithm's *effect*. Navigation on roads is mostly premised on rules-based behaviour that require the knowledge of traffic rules and the ability to interpret and react to uncertainty. In AVs, among the tasks being automated is the identification and classification of objects usually encountered while driving, such as other vehicles, traffic signs, traffic lights and road lining.[20] Furthermore, 'situational awareness and interpretation' is also being automated. For example, AVs should be able 'to distinguish between ordinary pedestrians (merely to be avoided) and police officers giving direction' and conform to social habits and rules by, for example, 'interpret[ing] gestures by or eye contact with human traffic participants'.[21] Finally, there is an element of prediction: AVs should have the capability to anticipate the behaviour of human traffic participants.[22]

In the design of AVs, the question of whether traffic rules can be accurately embedded in algorithms and, if so, who is concretely responsible for translating these rules into algorithms – for instance, is it only programmers or are lawyers and/or manufactures also involved? – becomes relevant to determine the accuracy of the algorithm design as well as a possible distribution of criminal responsibility. In this respect, it must be taken into consideration that while some traffic rules are relatively precise and consist of specific obligations (e.g., a speed limit represents an obligation *not* to exceed that speed),[23] there are also several open textured and context-dependent traffic norms (e.g., regulations requiring drivers to drive carefully/prevent danger on the road).[24]

Turning to AV incidents, these might stem from a failure of the AI to identify objects or correctly classify them. For example, the first widely reported incident involving an AV, which occurred in May 2016, was allegedly caused by the vehicle's sensor system failing to distinguish a large white truck crossing the road from the bright spring sky.[25] Incidents may

[19] See sections 4 and 5.

[20] H. Prakken, 'On the Problem of Making Autonomous Vehicles Conform to Traffic Law', 25 *Artificial Intelligence and Law* 341 (2017), 353.

[21] *Ibid.*, 354.

[22] *Ibid.*, 354.

[23] *See* Prakken's analysis of Dutch traffic laws which could be extended to other similar European systems by analogy, *supra* note 20, 345, 346 and 360. However, Prakken also provides an overview of open textured and vague norms in Dutch traffic law, 347 and 348.

[24] Prakken, *ibid.*, 347 and 348. *See* open-textured traffic rules in the Swiss Traffic Code Arts. 4, 26, 31 Straßenverkehrsgesetz (StVG).

[25] D. Yadron and D. Tynan, 'Tesla driver dies in first fatal crash while using autopilot mode', *The Guardian (online)*, 1 July 2016, available at: www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk. *See* also where Tesla's Autopilot system did not detect a truck ahead in the road: J. Plungis, 'Tesla Driver in Fatal March Crash Was Using Autopilot, NTSB Says', *Consumer Reports*, 16 May 2019, available at: www.consumerreports.org/car-safety/tesla-driver-in-fatal-march-crash-was-using-autopilot-ntsb-says/. *See* also a fatal accident where Tesla's Autopilot function failed to notice 'that the road was ending and then drove past a stop sign and a flashing red light' available at: N.E. Boudette, 'Inside a Fatal Tesla Autopilot

also arise due to failures to correctly interpret or predict the behaviour of others and traffic conditions (which may sometimes be interlinked with or compounded by problems of detection and sensing).[26] In turn, mistakes in both object identification and prediction might occur as a result of faulty algorithm design and/or derive from the data. In the former case, *prima vista*, if mistakes in object identification and/or prediction occur due to an inadequate algorithm design, the criminal responsibility of the programmer(s) could be engaged.

In relation to the latter, the increasing and almost dominant use of machine learning (ML) algorithms in AVs[27] make the issue of algorithms and data interrelated, with the performance of algorithms becoming heavily dependent on the quality of data. A multitude of different algorithms are used in AVs for different purposes, with supervised and unsupervised learning-based algorithms often complementing one other. Supervised learning is where an algorithm is fed instructions on how to interpret the input data. As such, supervised learning relies on a fully labelled dataset. Within AVs, the supervised learning models are usually: 1) 'classification' or 'pattern recognition algorithms', which process a given set of data into classes and help to recognise categories of objects in real time, such as street signs; and, 2) 'regression', which is usually employed for predicting events.[28] In cases of supervised learning, mistakes can arise due to incorrect data annotation instead of a faulty algorithm design *per se*. If incidents *do* occur, for example, due to wrong data labelling carried out by a third party (e.g., data annotation companies),[29] arguably the programmer could not foresee those risks and be considered 'in *control*' of the subsequent navigation decisions.

Other issues may arise with unsupervised learning[30] where a ML algorithm receives unlabelled data and programmers 'describe the desired behaviour and teach the system to perform well and generalise to new environments through learning'.[31] Data can be provided in the phase of simulating and testing but also during the use itself by the end-user. Within such methods, 'deep learning' is increasingly used to improve navigation in AVs. Deep learning is a form of unsupervised learning that 'automatically extracts features and patterns from raw data [such us real time data] and makes predictions or takes actions based on some reward function'.[32] When an incident occurs due to deep learning techniques using real data, it must be assessed whether the programmer could foresee that specific risk and the resulting harm or whether it, for example, derived from an unforeseeable interaction with the environment.

Accident: It Happened So Fast', *The New York Times*, 17 August 2021, available at: www.nytimes.com/2021/08/17/business/tesla-autopilot-accident.html.

[26] *See*, for example, the accident involving a Tesla Model 3 which hit a Ford Explorer pickup truck, killing one passenger: N.E. Boudette, 'Tesla Says Autopilot Makes Its Cars Safer. Crash Victims Say It Kills', *The New York Times*, 5 July 2021, available at: www.nytimes.com/2021/07/05/business/tesla-autopilot-lawsuits-safety.html.

[27] upGrad, 'How Machine Learning Algorithms Made Self Driving Cars Possible?', *upGrad Blog*, 18 November 2019, available at: www.upgrad.com/blog/how-machine-learning-algorithms-made-self-driving-cars-possible/.

[28] *See* Mindy Support, 'How Machine Learning in Automotive Makes Self-Driving Cars a Reality', *Mindy News Blog*, 12 February 2020, available at: mindy-support.com/news-post/how-machine-learning-in-automotive-makes-self-driving-cars-a-reality/.

[29] *See ibid.*

[30] Unsupervised models include: 'clustering' which is used, for example, when the (supervised) classification algorithms have failed to identify an object; 'simulation' and 'test data generation' which are used for building 'synthetic data' to virtually train algorithms and increase the efficiency of driverless cars in unpredictable conditions; 'anomaly detection', which is used for trying to recognise abnormal behaviours of the autonomous vehicle itself and the operator. *See* V. Haydin, 'What does Unsupervised Learning Have in Store for Self-Driving Cars'?, *intellias*, available at: intellias.com/what-does-unsupervised-learning-have-in-store-for-self-driving-cars/.

[31] S. Kuuti *et al*, 'A Survey of Deep Learning Applications to Autonomous Vehicle Control', 22 *IEEE Transactions on Intelligent Transportation Systems* 1 (2021).

[32] B. Gupta, A. Anpalagan, L. Guan, and A.S. Khwaja, 'Deep Learning for Object Detection and Scene Perception in Self-Driving Cars: Survey, Challenges, and Open Issues', 10 *Array* 1 (2021), 8.

2.2 Programmer or User: Who is in Control of AVs?

As shown in the March 2018 Uber incident,[33] incidents can also derive from failures of the user to regain control of the AV. In these situations, some AV manufacturers attempt to shift the responsibility for ultimately failing to avoid collisions onto the AVs' occupants.[34] However, there are serious concerns as to whether an AV's occupant – who is essentially in an oversight role, depending on the level of automation – is cognitively in the position to regain control of the vehicle. This is also known as automation bias,[35] a cognitive phenomenon which occurs in human-machine interaction, where complacency, decrease of attention, and overreliance on the technology might impair humans from overseeing, intervening and overriding the system if needed.

Faulty human-machine interface (HMI) design – the technology which connects an autonomous system to the human, such as a dashboard or interface – could cause the inaction of the driver in the first place. In these instances, the driver could be relieved from criminal responsibility. Arguably, HMIs do not belong to programmers' functional obligations either and fall beyond a programmer's *control*.

There are phases other than actual driving where a user could gain *control* of an AV's decisions. Introducing ethics settings into the design of AVs may ensure control over a range of morally significant outcomes, including trolley-problem-like decisions.[36] Such settings may be mandatory (i.e. introduced by manufacturers with no possibility from users to intervene and/or customise them) or customisable by users.[37] Customisable ethics settings allow users 'to manage different forms of failure by making autonomous vehicles follow [their] decisions' and their intention.[38] Where introduced, customisable ethics settings transfer the *control* over algorithmic decision-making from programmers to users and, any consequent responsibility for incidents.

2.3 Are there AV- Related Risks 'Out of' Programmers' Control?

There are a group of risks and failures that could be considered outside the *control* of programmers, that being communications failures, hacking of the AV by outside parties, and unforeseeable bystander behaviour. One of the next predicted steps in vehicle automation is the development of software enabling AVs to communicate with each other and to share real-

[33] *See* Levin and Wong, *supra* note 1.

[34] *See* Grieman, *supra* note 6.

[35] K.L. Mosier and L.J. Skitka, 'Human Decision Makers and Automated Decision Aids: Made for Each Other?' in R. Parasuraman and M. Mouloua (eds.), *Automation and Human Performance: Theory and Applications* (CRC Press: 1996), 201.

[36] 'Trolley-problem-like scenarios refer to situations where all available options lead to different forms of costly failures' such as situations where 'user-vehicle system needs to either steer the system to the left and potentially fall from a bridge, steer to the right and run over some cyclists, or go straight forward and hit pedestrians who have just stepped onto the road without properly scanning for oncoming vehicles', S. Soltanzadeh, J. Galliott, and N. Jevglevskaja, 'Customizable Ethics Settings for Building Resilience and Narrowing the Responsibility Gap: Case Studies in the Socio-Ethical Engineering of Autonomous Systems', 26 *Science and Engineering Ethics* 2693 (2020), 2696.

[37] *Ibid.*, 2705.

[38] *Ibid.*, 2697.

time data gathered from their sensors and computer systems.[39] This ultimately means that a single AV 'will no longer make decisions based on information from just its own sensors and cameras, but it will also have information from other cars'.[40] Failures in vehicular communication technologies[41] or inaccurate data collected by other AVs cannot be attributed to a single programmer as they might fall beyond their responsibilities and functions (thus also their *control*).

Hacking could also cause AV incidents. For example, it has been showed that 'placing stickers on traffic signs and street surfaces can cause self-driving cars to ignore speed restrictions and swerve headlong into oncoming traffic'.[42] Here the criminal responsibility of a programmer could depend on whether the attack could have been anticipated and whether the programmer should have created safe-guards against it. However, the complexity of AI systems could make them more difficult to defend from attacks and more vulnerable to adversarial interference.[43]

Finally, imagine an AV that, while correctly following traffic rules, hits a pedestrian who had unforeseeably slipped and fallen onto the road. Such, unforeseeable behaviour of a bystander is relevant in criminal law cases on vehicular homicide as it will break the causal nexus between the programmer and the harmful outcome.[44] In the present case, it must be determined which unusual behaviour should be anticipated at the stage of programming and whether standards of anticipation in AVs should be higher than for humans.

3. Risks Posed by AWs and Programmers' Control

While again not intending to provide a comprehensive overview, Section 3 follows the structure of Section 2 in addressing some of the risks inherent in AWs – including algorithms, data, users, communication technology, hacking and adversarial interference, and the unforeseeable behaviour of individuals in war – and distinguishing them on the basis of their causes and programmers' level of *control* over them. While some risks cannot be predicted, the 'development of the weapon, the testing and legal review of that weapon, and th[e] system's previous track record' could provide information about the risks involved in the deployment of AWs.[45] Some risks could thus be understood and foreseen by the programmer and therefore be considered under their *control*.

3.1 Are Programmers in Control of Algorithm and Data-Related Risks in AWs?

[39] K. Harel, 'Self-driving cars must be able to communicate with each other', *Aarhus University Department of Electrical and Computer Engineering: News*, 2 June 2021, available at: ece.au.dk/en/currently/news/show/artikel/self-driving-cars-must-be-able-to-communicate-with-each-other/.

[40] Harel, *ibid*.

[41] *See* on this topic, M.N., Ahangar, Q.Z. Ahmed, F.A. Khan, and M. Hafeez, 'A Survey of Autonomous Vehicles: Enabling Communication Technologies and Challenges', 21 *Sensors* 706 (2021).

[42] K.J. Hayward and M.M. Maas, 'Artificial Intelligence and Crime: A Primer for Criminologists', 17 *Crime Media Culture* 209 (2021), 216.

[43] M. Caldwell, J.T.A. Andrews, T. Tanay, and L.D. Griffin, 'AI-Enabled Future Crime', 9 *Crime Science* 14 (2020) 22.

[44] *See* section 4.

[45] M.A. Holland, *Known Unknowns: Data Issues and Military Autonomous Systems*, (UNIDIR: 2021), 10.

In this sub-section, I will take autonomous drones as an example of one of the most likely applications of autonomy within the military domain[46] to highlight the tasks increasingly autonomous in AWs, materialising in the *behaviour* and *effects* of AWs; the rules to be programmed; and identify where risks might lie in the phase of algorithm design.

Within autonomous drones, two of main tasks which are automated are: 1) navigation, which is less problematic than on roads and a relatively straightforward rule-based behaviour (e.g., they simply must avoid obstacles while in flight); and, 2) weapon release, which is much more complex as 'ambiguity and uncertainty are high when it comes to the use of force and weapon release, bringing this task in the realm of expertise-based behaviours'.[47] Within the weapon release function, target identification is the most important function since it is crucial to ensure compliance with the international humanitarian law (IHL) principle of distinction, the violation of which could also give rise to individual criminal responsibility for war crimes. The principle of distinction holds that belligerents and those executing attacks must distinguish at all times between civilians and combatants (and therefore civilians must not be targeted).[48] In target identification, the main two tasks that are automated are: 1) object identification and classification on the basis of pattern recognition;[49] and 2) prediction (for example predicting that someone is surrendering or, based on the analysis of patterns of behaviour, predicting that someone is a lawful target).[50]

Some of the problems that may arise in the algorithm design phase derive from translating rules of IHL,[51] such as the principle of distinction, into algorithms, as well as programming into code knowledge and expert-based rules,[52] such as those that are needed to analyse patterns of behaviour in targeted strikes. These legal concepts are open textured and context-dependent [53] This phase presents some differences when compared to an AV context. Arguably traffic law is more widely understood by programmers than the relatively niche and context-specific nature of IHL. As will be highlighted below, programming IHL notions requires a stronger collaboration with outside expertise – namely, military lawyers and operators – and thus a possible distribution of responsibility.

Instead, similar observations to those made above in relation to supervised and unsupervised learning algorithms can be made regarding the responsibility of AW programmers. *Prima vista*, if harm results from mistakes in object identification and prediction that occur due to an inadequate algorithm design, the criminal responsibility of the programmer(s) could be engaged. However, depending on the foreseeability of such data failures to the programmer and the involvement of third parties in data labelling (whose

[46] M. Ekelhof and G.P. Paoli, *Swarm Robotics: Technical and Operational Overview of the Next Generation of Autonomous Systems* (UNIDIR: 2020), 51.

[47] A.A. Melancon, 'What's Wrong with Drones? Automatization and Target Selection', 31 *Small Wars and Insurgencies* 801 (2020), 806.

[48] The principle of distinction is enshrined in Art. 48 of AP I with accompanying rules in Arts 51 and 52 of AP I.

[49] A. Deeks, 'Coding the Law of Armed Conflict: First Steps', 49 *University of Virginia School of Law: Public Law and Legal Theory Paper Series* (2020), 4 and 5: 'the military might seek to classify a different set of objects: people holding weapons in a hostile pose. Once the algorithm identifies a threat (weapon or person) to some pre-determined level of certainty, the military unit deploying the algorithm will choose how to respond'; Melancon, *supra* note 47, 12 and 13.

[50] Let us think, for example, of autonomous drones equipped with autonomous or automatic target recognition (ATR) software to be employed for targeted killings of alleged terrorists or 'Project Maven' which entails the use of big data and machine learning in order to automate the work of analysts assessing drone-collected video surveillance footage and whose analysis is used to support militaries in target selection.

[51] On the challenges, *see* A. Schuller, 'Artificial Intelligence Effecting Human Decisions to Kill: The Challenge of Linking Numerically Quantifiable Goals to IHL Compliance', 15 *ISJLP* 105 (2019).

[52] Melancon, *supra* note 47, 14-16.

[53] Deeks, *supra* note 49, 10.

mistakes cannot be anticipated), criminal responsibility might not be attributable to programmers. Similar to AVs, the increasing use of deep learning methods in AWs make algorithms' performance dependent on both the availability and accuracy of data. Low quality and incorrect data, missing data, and/or discrepancies between real and training data may be conducive to the mis-identification of targets.[54] When unsupervised learning is used, environmental conditions and armed conflict-related conditions (e.g., smoke, camouflage and concealment) may inhibit the collection of accurate data. [55] In the case of supervised learning, errors in data may, instead, lie in 'human-generated data feed'[56] and the incorrect labelling of data could lead to mistakes and incidents that might not be criminally attributable to programmers.

3.2 Programmer or User: Who is in Control of AWs?

The relationship between programmers and users of AWs presents some different challenges when compared with AVs. In light of current trends in the development of AW – arguably towards human-machine interaction rather than full autonomy of the weapons system – the debate has focused on the degree of *control* that militaries must retain over the weapon release functions of AWs.[57]

However, *control* can be shared between and distributed among programmers and users in different phases, spanning from design to deployment. In fact, AI engineering in the military domain might require a very strong collaboration between programmers and military lawyers in order to accurately code IHL rules in algorithms.[58] Those arguing for the (albeit debated) introduction of ethics settings in AWs argue that ethics settings would 'enable humans to exert more control over the outcomes of weapon use [and] make the distribution of responsibilities [between manufacturers and users] more transparent'.[59]

Finally, given their complexity, programmers of AWs might be involved more than programmers of AVs in the *use* of AWs and in the targeting process.[60] In these situations, it must be evaluated to what extent a programmer could anticipate and foresee a certain risk entailed in the deployment and *use* of an AW in relation to a specific attack rather than just its use in the abstract.

3.3 Are there AW-Related Risks 'Out of' Programmers' Control?

It is highly likely that AWs will be subject to adversarial interference by enemy forces. An UNIDIR report lists several pertinent examples: a) signal jamming could 'block systems from receiving certain data inputs (especially navigation data)'; b) hacking, such as 'spoofing' attacks, might 'replace an autonomous system's real incoming data feed with a fake feed containing incorrect or false data'; c) 'input' attacks could 'change a sensed object or data

[54] *See* Holland, *supra* note 45, 4; J. Hughes, 'The Law of Armed Conflict Issues Created by Programming Automatic Target Recognition Systems Using Deep Learning Methods', 21 *YIHL* 99 (2018), 106 and 107.

[55] Holland, *supra* note 45, 6.

[56] Holland, *ibid.*, 4.

[57] In debates over AWs, the issue of (meaningful human) control has been primarily discussed with respect to the military operators/users or in terms of comprehensive human oversight or control over the lifecycle of the weapon, but not to the same extent regarding the pre-duse phase.

[58] Deeks, *supra* note 49, 11.

[59] Soltanzadeh et al, *supra* note 36, 2704 and 2705.

[60] Military targeting must be intended as encompassing more than critical functions of weapon release.

source in such a way as to generate a failure', for example, enemy forces 'may seek to confound an autonomous system by disguising a target; and, d) 'adversarial examples' or 'evasion' which are attacks that 'involve adding subtle artefacts to an input datum that result in catastrophic interpretation error by the machine' might occur.[61] In such situations, the issue of criminal responsibility for programmers will depend on the modalities of the adversarial interference, whether it could have been anticipated, and whether the AW could have been protected from foreseeable types of attacks.

Similar to the AV context, failures of communication technology – caused by signal jamming or by failures of communication systems themselves – between a human operator and the AI system or among AI systems themselves (such as within swarms of drones) may lead to incidents that could not be imputed to a programmer.

Finally, conflict environments are likely to drift constantly as '[g]roups engage in unpredictable behaviour to deceive or surprise the adversary and continually adjust (and sometimes radically overhaul) their tactics and strategies to gain an edge'.[62] The continuously changing and unforeseeable behaviour of opposing belligerents and the tactics of enemy forces can lead to 'data drift', whereby changes that are difficult to foresee can lead to a weapon system's failure without it being imputable to a programmer.

4. AV-Related Crimes on the Road and AW-Related War Crimes on the Battlefield

The following paragraphs will distil the legal ingredients of crimes against persons resulting from failures in the use of AVs and AWs. The key question therein analysed is whether the *actus reus*, i.e. the prohibited conduct including its resulting harm, could ever be carried out by programmers of AVs and AWs. From the analysis that follows, it emerges that save for war crimes under the Rome Statute, which prohibit a conduct, the crimes on the road and the battlefield currently under examination are formulated as result crimes – they require the causation of harm (e.g. death or injuries). In relation to crimes of conduct, the central question is whether programmers *controlled* the *behaviour* of an AV and AWs: e.g. the AWS's direction of an indiscriminate attack against civilians. In relation to crimes of result, the central question is whether programmers exercise *causal control* over a chain of events leading to death, i.e. over the *behaviour* and the *effects* of AVs and AWs. While arguably the establishment of causation with regard to crimes of conduct and result might raise different issues in light of the causal gap that characterise the latter, crimes committed with the intermediation of AI raise similar problems in terms of causation.[63] Crimes committed with the intermediation of AI, be they of conduct or result, present, respectively, a causal gap between a programmers' conduct and,the unlawful behavior or effect of an AV and AW. Thus the issue is causal nexus between a programmers' conduct and either the *behaviour* (in the case of crimes of conduct), or the *effects* (in the case of crimes of result) *of* AVs and AWs. To illustrate these issues, sections 4.1 and 4.2 will describe the *actus reus* of AV- and AW-related crimes while section 4.3 will turn the question of *causation*. While the central question of this chapter concerns the *actus reus*, at

[61] Holland, *supra* note 45, 7.

[62] *Ibid.*, 9.

[63] Alleged differences might lie in the fact that crimes of conduct 'rest on an immediate connection between the harmful action and the relevant harm' and that crimes of result 'are characterized by a [special and temporal] causal gap between action and consequence', Fletcher, *Basic Concepts of Criminal Law*, OUP (1998), 61.

the end of this section, I will also make some remarks on *mens rea* and the relevance of risk-taking and negligence in this debate.

4.1 Actus Reus *in AV-Related Crimes*

This sub-section is concerned with the domestic criminal offences of negligent homicide and manslaughter with the purposes of assessing whether the *actus reus* of AV-related crimes could be performed by a programmer. It does not address traffic and road violations generally,[64] nor the specific offence of vehicular homicide.[65]

Given the increasing use of AVs and pending AV-related criminal cases in the United States (US),[66] it seems appropriate to take the Model Penal Code (MPC) as an example of common law legislation.[67] According to the MPC the *actus reus* of manslaughter consists of 'killing for which the person is reckless about *causing* death'.[68] Negligent homicide concerns instances where a 'person is not aware of a substantial risk that a death will *result* from his or her conduct, but should have been aware of such a risk'.[69]

Taking Germany as a representative example of civil law traditions, the German Criminal Code (StGB) distinguishes two forms of intentional homicide: murder[70] and manslaughter.[71] Willingly taking the risk of *causing* death is sufficient for manslaughter.[72] Negligent homicide is proscribed separately[73] and the *actus reus* consists of *causing* the death of a person through negligence.[74]

These are crimes of result where the harm consists of the death of a person. While programmers' conducts may be remote with respect to incidents with AVs, some decisions taken by programmers at an early stage of development of an AV could have a decisive impact on navigation *behavior* of an AV possibly resulting in death. In other words, it is conceivable that a faulty algorithm designed by a programmer could *cause* a fatal road accident. The question is thus the threshold of *causal control* exercised by programmers over the unlawful *behaviour* (navigation) and its unlawful *effects* (death) of an AV.

4.2 Actus Reus *in AW-Related War Crimes*

This sub-section is concerned with AW-related war crimes in order to assess whether programmers could fulfil their *actus reus*. Since they will most likely stem from AWs'

[64] *See* on this topic H. Prakken, *supra* note 20.

[65] While the US' Model Penal Code (MPC) does not contain a provision dealing with vehicular homicide, legislations in certain domestic systems envisage it.

[66] *See supra* note 4.

[67] American Law Institute, *Model Penal Code: Official Draft and Explanatory Notes: Complete Text of Model Penal Code as Adopted at the 1962 Annual Meeting of the American Law Institute at Washington, D.C., May 24, 1962* (American Law Institute: 1985).

[68] MPC, §2.13(1)(b). See P.H. Robinson, 'United States' in *The Handbook of Comparative Criminal Law*, *supra* note 13**Error! Bookmark not defined.**, 585.

[69] Robinson, *ibid.* (emphasis added).

[70] Strafgesetzbuch (StGB), Criminal Code in the version published on 13 November 1998 (Federal Law Gazette I, p. 3322), as last amended by Article 2 of the Act of 19 June 2019 (Federal Law Gazette I, p. 844), §211(1) (emphasis added).

[71] Under German criminal law, manslaughter is the intentional killing of another person without aggravating circumstances, StGB, §212.

[72] T. Weigend, *supra* note 13, 262.

[73] StGB, §222.

[74] Weigend, *supra* note 13, 263 (emphasis added).

incapability to distinguish between civilian and military targets, the war crime of indiscriminate attacks against civilians, criminalising violations of the aforementioned IHL rule of distinction,[75] becomes of crucial relevance.[76]

The war crime of indiscriminate attacks refers to: a) attacks with weapons which are *not inherently indiscriminate*, but that are *used in an indiscriminate manner* against civilians; or b) attacks *using inherently indiscriminate* weapons, i.e., weapons incapable of distinguishing between civilian objects and military objectives.[77] Instances where programmers are involved in the *indiscriminate use* of an AW are possible but less likely. It is rather the latter scenario and their possible role in programming AWs that are *inherently indiscriminate* i.e., incapable of differentiating between lawful and unlawful targets that could trigger programmers' criminal responsibility as principals.

It must be noted that this war crime is neither specifically codified in the Rome Statute nor in AP I, but has been subsumed by international criminal courts[78] under the war crime of directing attacks against civilians.[79] The war crime provision in API is a result crime, i.e., the *actus reus* is defined in terms of causing death or injury. When a criminal provision is characterised by the causation of a harmful result that must occur in addition to the conduct, a causal nexus between the *effects* resulting from the deployment of an AW a programmer's conduct must be established. In the Rome Statute, the war crime is formulated as a conduct crime, proscribing as *actus reus* the 'directing of an attack' against civilians.[80] Thus a causal nexus must be established between the unlawful AW's *behaviour* and the programmer's conduct. Moreover, conduct crimes could entail a result and in this particular case has also been interpreted as entailing 'an attack' as result– rather than deaths and/or injuries resulting from the attack. In both war crimes, thus the question of *causal control* exercised by programmers over the *behaviour* and/or *effects* (death or attack) of an AW.

A final issue relates to the temporal and geographical applicability of the law of war crimes which may be a challenge to the ascription of programmers' criminal responsibility. The law of war crimes applies from the initiation to the end of an armed conflict and some war crimes must take place 'in the context of and was associated with' it.[81] Due to the temporal and physical distance between the programmers' conducts and the armed conflict, this threshold may be difficult to fulfil. However, first, it is conceivable that programmers program AWS software or upgrade them during the armed conflict. Second, in line with this Chapter's thesis, one could argue that programmers' control continues even after the completion of the 'act of programming' and the effects of decisions taken by programmers materialise in the *behaviour*

[75] For the underlying IHL, *see* Article 51(4)(a) of AP I; *see* also ICRC, Customary International Humanitarian Law Study, Rule 12,

[76] *See* M. Bo, 'Autonomous Weapons and the Responsibility Gap in light of the *Mens Rea* of the War Crime of Attacking Civilians in the ICC Statute', 19 *JICJ* 275 (2021), 282-285.

[77] K. Dörmann, *Elements of War Crimes under the Rome Statute of the International Criminal Court: Sources and Commentary* (CUP: 2003), 131 and 132.

[78] Both by the ICC and the International Criminal Tribunal for the former Yugoslavia (ICTY). The latter interpreted violations of Article 3 of its Statute, relevant to unlawful attack charges, by resorting to Article 85(3) of AP I, *see* Bo, *supra* note 76, 283 and 284.

[79] Article 8(2)(b)(i) establishes that, in the context of an international armed conflict (IAC), intentionally directing 'attacks against the civilian population as such or against individual civilians not taking direct part in hostilities' is a war crime under the jurisdiction of the ICC. The same war crime is listed under Article 8(2)(e)(i) in relation to non-international armed conflicts (NIAC). Article 85(3) of AP I, the *actus reus* of the war crime of wilfully launching attacks against civilians (as well as the other grave breaches enshrined in this Article) contains the requirement that an attack against civilians causes 'death or serious injury to body or health'.

[80] A. Eser, 'Mental Elements – Mistake of Fact and Mistake of Law', in A. Cassese, P. Gaeta, and J.R.W.D. Jones (eds.), *The Rome Statute of the International Criminal Court: A Commentary* (OUP: 2002), 911.

[81] Element 4 of the elements of the crime at Article 8(2)(b)(i) of the Rome Statute

and/or *effects* of AWs in armed conflict. Thus, the programmers exercise a form of *control* over the *behaviour* and/or *effects* of AWs that begins with the act of programming and continues after.

4.3 *The Causal Nexus between Programming and AV- and AW-Related Crimes*

Crucially for discussing programmers' criminal responsibility is the *causal control* exercised by programmers over the behaviour and/or effects of AVs and AWs. The assessment of causation refers to the conditions under which such an AV- and AW's unlawful *behaviour* and/or *effects* should be deemed the 'result' of programmers' conduct for the purpose of holding them criminally responsible.

Causality is a complex topic. In common law and civil law countries, several tests to establish causation in legal terms have been put forward. Due to difficulties in establishing a uniform test for causation, it has been argued that determining conditions for causation are 'ultimately a matter of legal policy'.[82] However, this does not mean the formulation of causality tests aimed at achieving policy and objectives pursued by the relevant criminal provisions' is not important. While a comprehensive analysis of these theories is beyond of the scope of the present chapter, for the purposes of establishing when programmers' exercise *causal control* some theories including elements of foreseeability are relevant and arguably in line with the policy objectives pursued in the context of the repression AV- and AW-related crimes.

First, in common law and civil law countries the 'but-for'/*conditio sine qua not* test is the dominant test for establishing physical causation intended as a relationship of physical cause-effect.[83] In the language of MPC §2.03(1)(a), the conduct must be 'an antecedent but for which the result in question would not have occurred'. The 'but for' test works satisfactorily in cases of straightforward cause and effect (pointing a loaded gun towards the chest of another person and pulling the trigger). However, AV- and AW-related crimes are characterised by a temporal and physical gap between programmers' conduct and the behaviour and effect of AVs and AWs. They involve complex interactions between AVs and AWs, on the one hand, and humans (programmers, data providers and labellers, users, etc.), on the other hand. What is more AI itself is a factor that could intervene in the causal chain. The problem of causation in these cases must thus be framed considering the relevance of intervening and superseding causal forces which may break the causal nexus between a programmer's conduct and an AV- and AW-related crime.

Both civil law and common law systems have adopted several theories to overcome some of the shortcomings[84] and correct the potential over-inclusiveness[85] of the 'but-for' test in complex cases involving numerous necessary conditions. From the outset it is important to note that some of these theories include elements of foreseeability in the causality test.

The MPC adopts the 'proximate cause test' which 'differentiates among the many possible "but for" causal forces, identifying some as "necessary conditions" – necessary for the result to occur but not its direct 'cause' – and recognising others as the "direct" or "proximate"

[82] Ryu, *supra* note 17, see contra Fletcher highlighting the importance of the principle of legality in criminal, law, *supra* note 63, 66.

[83] *See* Ryu, *ibid*, 787. Also described as 'empirical causality', which refers to the 'metaphysical [and deterministic] question of cause and effect', M. Cupido, 'Causation in International Crimes Cases: (Re)Conceptualizing the Causal Linkage', 32 *Criminal Law Forum* 1 (2021), 24.

[84] Ryu, *ibid.*

[85] Ryu, *ibid.*

cause of the result'.[86] The relationship is 'direct' when the result is foreseeable and as such 'this theory introduces an element of culpability into the law of causation'.[87]

German theories about adequacy held that whether a certain factor can be considered a cause of a certain effect depends on 'whether conditions of that type do, generally, in the light of experience, produce effects of that nature'.[88] These theories, which are not applied in their pure form in any criminal law, include assessments that resemble a culpability assessment. They, in fact, bring elements of foreseeability (and thus culpability) into the causality test and, in particular, a probability and possibility judgement on the part the accused.[89] However, these theories leave unresolved the different knowledge perspectives, i.e. objective, subjective or mixed, on which the foreseeability assessment is to be based.[90]

Other causation theories entail an element of understandability, awareness, and anticipation of risks. In the MPC, the 'harm-within-the risk' theory considers that causation in reckless and negligent crimes is in principle established when the result was within the 'risk of which the actor is aware or […] of which he should be aware'.[91] In German criminal law, some theories describe causation in terms of the creation or aggravation of risk and limit causation to the unlawful risks that the violated criminal law provision intended to prevent.[92]

In response to the drawbacks of these theories, the teleological theory of causation holds that in all cases involving a so-called intervening independent causal force, the criterion should be whether the intervening causal force was 'produced by "chance" or was rather imputable to the criminal act in issue'.[93] Someone would be responsible for the result if their act contributed in any manner to the intervening factor. What matters is the accused's *control* over the criminal conduct and whether the intervening factor was connected in a but/for sense to their criminal act,[94] thus falling within their control.

In international criminal law, a conceptualisation of causation that goes beyond the physical relation between acts and effects is more embryonic. However, it has been suggested that theories drawn from national criminal law systems, such as risk-taking and linking causation to culpability, thus to foreseeability, should inform a theory of causation in international criminal law.[95] Importantly, it has been suggested that causality should entail an evaluation of the 'functional obligations' of an actor and their area of operation in the economic sphere. In this context, causation is 'connected to an individual's *control* and scope of influence' and is limited to 'dangers that he creates through his activity and has the power to avoid'.[96] Upheld in the context of international crimes, which have a collective dimension,

[86] A. Leavens, 'A Causation Approach to Criminal Omissions', 76 *California Law Review* 547 (1988), 564.

[87] Ryu, *supra* note 17, 789.

[88] Ryu, *ibid.* 791.

[89] Ryu, *ibid.*, 792.

[90] Ryu, *ibid.* 795.

[91] MPC, §2.03 (3). MPC, §2.03(2) and (3) formulate several exceptions to the general proximity standard in cases of intervening and superseding causal forces.

[92] Among the 'but-for' conditions that are *not* considered attributable are: '[a] consequence that the perpetrator has caused […] if that act did not unjustifiably increase a risk'; '[a] consequence was not one to be averted by the rule the perpetrator violated'; and 'if a voluntary act of risk taking on the part of the victim or a third person intervened'. Weigend provides the three following examples respectively: 'a driver hits a pedestrian when the driver has followed the traffic rules and the pedestrian has unforeseeably slipped and has fallen onto the street'; the driver had been speeding five minutes before the incident; had he stayed within the permissible speed limit, he would not have been at the place when the pedestrian slipped'; 'the pedestrian threw himself before the speeding driver's car in order to commit suicide', Weigend, *supra* note 13**Error! Bookmark not defined.**, 268. See also Cupido, *supra* note 83, 26 and 27.

[93] Ryu, *supra* note 17, 797.

[94] Ryu, *ibid.*, 798.

[95] Cupido, *supra* note 83, 43-47.

[96] Cupido, *supra* note 83, 41.

these theories could usefully be employed in the context AV and AW development, which equally has a collective nature and is characterised by a distribution of responsibilities.

Importantly, programmers in some instances will cause harm through omission, notably where programmers fail to avert a particular harmful risk when they are under a legal duty to prevent harmful events of that type ('commission by omission') [97] In these cases, the establishment of causation will be hypothetical as there is no physical cause-effect relationship between an omission and the proscribed result.[98] Other instances concern whether negligence on the side of the programmers – such as a lack of instructions and warnings – has contributed and caused the omission (failure to intervene) on the part of user. Such omissions amount to negligence, i.e violations of positive duties of care,[99] and since it belongs to *mens rea* will be addressed in the following sub-section.

4.3 Criminal Negligence: Programming AVs and AWs

An assessment of programmers' criminal responsibility would be incomplete without addressing *mens rea* issues, also in light of the aforementioned inclusion of culpability assessments into causation tests. In relation to the *mens rea*, while intentionally and knowingly programming an AV or AW to commit crimes falls squarely under these prohibitions, both in an AV and AW context the most expected and problematic issue is the *unintended* commission of these crimes, i.e., cases in which the programmer did not design the AI system to commit an offence, but harm nevertheless arises during its use.[100] In such situations, programmers had no intention to commit an offence, but still risked criminal liability. To define the scope of criminal responsibility for unintended harm is crucial to determine which risks can be known and foreseeable by an AV or AW programmer.

There are important differences in the *mens rea* requirements of the crimes under scrutiny, the most important being that the offence of negligent homicide under domestic law (which might apply to programmers of AVs) does not have a parallel 'negligent attacks against the civilian population' war crime for which programmers of AWs could be held responsible under.

It is beyond dispute that under domestic criminal law, the standards or recklessness and negligence apply to the AV-related crimes of manslaughter and negligent homicide. The crucial distinction between recklessness and negligence is that: '[a] person acts "recklessly" with respect to a result if he or she *consciously disregards a substantial risk* that his or her conduct will cause the result; [whereas] he or she acts only "negligently" if he or she is *unaware of the substantial risk but should have perceived it*'.[101] The MPC provides that 'criminal homicide

[97] StGB, §13 provides that: '[w]hoever omits to avert a consequence that is part of a statutory offense description is punishable according to that statute only if he is legally responsible for averting that consequence and when the omission is equivalent to actively committing the offense'.

[98] On causation in criminal omissions, *see* G. Hughes, 'Criminal Omissions', 67 *Yale LJ* (1958) 590, 627-631. Causation in 'commission by omission' is strictly connected with duties to act and duty to prevent a certain harm, *see* G.P Fletcher, *Rethinking Criminal Law* (OUP: 2000), 606; *see* also the theory developed by Leavens whereby the essence of criminal omission liability 'lies in deciding which omissions can fairly be said to cause prohibited harms', Leavens, *supra* note 86, 562.

[99] See forthcoming work of this author in the *Journal of International Criminal Justice* (Issue 1, 2023).

[100] *See* also S. Gless, E. Silverman and T. Weigend, 'If Robots Cause Harm, Who Is to Blame? Self-Driving Cars and Criminal Liability', 19 *New Crim. L. Rev.* 412 (2016), 425:

[101] Robinson, *supra* note 69, 575. *See* also Binder: ''Negligent manslaughter' now usually requires objective foreseeability of death, rather than the simple violation of a duty of care', G. Binder, 'Homicide on the Road' in M. Dubber and T. Hörnle (eds.), *The Oxford Handbook of Criminal Law*, (OUP: 2014), 710 (emphasis added).

constitutes manslaughter when it is committed recklessly',[102] meaning a 'killing for which the person is reckless about causing death and is reckless about the victim being a human being'.[103] Negligent homicide concerns instances where a 'person is not aware of a substantial risk that a death will result from his or her conduct, but should have been aware of such a risk'.[104] In the German criminal code, *dolus eventualis* (i.e., willingly taking the risk of causing death) would encompass situations covered by recklessness and is sufficient for manslaughter.[105] For negligent homicide[106] one of the prerequisites is that the perpetrator *can foresee the risk* for a protected interest.[107]

Risk-based *mentes reae* are more disputed in international criminal law. The International Tribunal for the former Yugoslavia (ICTY) accepted that recklessness could be a sufficient *mens rea* for the war crime of indiscriminate attacks under Article 85(3)(a) of AP I.[108] However, whether recklessness and *dolus eventualis* could be sufficient to ascribe criminal responsibility for war crimes within the framework of the Rome Statute remains debated.[109]

In sum, unlike incidents with AVs, incidents in war resulting from the negligence of a programmer cannot give rise to their criminal responsibility. Where applicable, recklessness and *dolus eventualis* – which entail understandability and foreseeability of risks of developing inherent indiscriminately AWs (with *dolus eventualis* requiring an additional 'acceptance component') – become crucial to ascribe responsibility to programmers in scenarios where programmers foresaw and took some risks. Excluding these mental elements would amount to ruling out the criminal responsibility of programmers in the most expected instances of war crimes.

5. Developing a(n) (International) Criminal Law-Infused Notion of *Meaningful Control* over AVs and AWs that Incorporates *Mens Rea* and Causation Requirements

This section elaborates on a notion of meaningful human control (MHC) applicable to AVs and AWs based on criminal law and that could function as a criminal responsibility 'anchor' or 'attractor'.[110]

First, it must be noted that this is not the first attempt to develop a conception of *control* applicable to both AVs and AWs. The notion of MHC developed by Santoni de Sio and Van den Hoven in the context of moral responsibility and AWs[111] has been recently extended to

[102] MPC, §2.13(1)(b).

[103] P.H. Robinson, *supra* note 69, 585.

[104] Robinson, *ibid*.

[105] T. Weigend, *supra* note 13, 262.

[106] StGB, §222

[107] There are four prerequisites for liability for criminal negligence: 'the actor can *foresee the risk* for a protected interest; the actor violates a duty of care with respect to the protected interest; harm as defined by the statute occurs; and the offender could have avoided the harm by careful conduct', Weigend, *supra* note 13, 263 (emphasis added).

[108] *See* the case law quoted in Bo, *supra* note 76, 293.

[109] Bo, *supra* note 76, 286-294.

[110] Amoroso and Tamburrini describe MHC as a responsibility attractor, D. Amoroso and G. Tamburrini, 'Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues', 1 *Current Robotics Reports* 187 (2020), 189.

[111] F.S. de Sio and J. van den Hoven, 'Meaningful Human Control over Autonomous Systems: A Philosophical Account', 5 *Frontiers in Robotics and AI* 1 (2018), 6-9.

AVs.[112] In their view, MHC should entail an element of traceability and trackability. Importantly, traceability entails that ‘*one human agent in the design history* or use context involved in designing, programming, operating and deploying the autonomous system […] *understands or is in the position to understand the possible effects* in the world of the use of this system’.[113] The requirement of traceability entails that someone in the *design* or *use* understands the capabilities of the AI system and its effects. Thus, programmers could exercise MHC.

In line with these studies, it is this chapter’s contention that programmers may decide and ‘control’ how both traffic law and IHL are embedded in the respective algorithms, how AI systems see and move, and how they react to changes in the environment. For Boutin, especially in the case of machine learning, programmers may have ‘genuine, although indirect, *control* over the AI system’.[114] Similarly, McFarland and McCormack affirm that programmers may exercise control not only over an abstract range, but also in relation to specific, behaviour and effects of AWs.[115] Against this background, this chapter contends that programmers’ *control* begins at the initial stage of the AI development process but continues in the use phase, extending to the *behaviour* and *effects* of AVs and AWs.

Second, based on the assumption of programmers’ control over certain AV and AW-related unlawful *behaviour* and *effects* how can MHC be conceptualised so to ensure that criminal responsibility is traced back to them? What are the elements that can be drawn from the above discussions on causality and *mens rea* in the context of AV- and AW-related crimes that could be relevant to MHC? From the foregoing discussion on causality, one can conclude that theories of causation that go beyond deterministic cause-and-effect assessments are particularly amenable to application to the present scenario. These theories either link causation to *mens rea* standards or describe it in terms of the aggravation of risk. In either case, the ability to understand the capabilities of AI systems and their effects, foreseeability, and anticipation of risks are required. Considering these theories of causation against recent studies on MHC over AVs and AWs, arguably the MHC’s requirement of traceability translates into the requirements of foreseeability and anticipation of risks.[116] In particular, due to the distribution of responsibilities in the context of AV and AW programming, causation theories that introduce the notion of function-related risks seem important to circumscribe programmers’ criminal responsibility to those risks within the respective obligation and thus sphere of influence and control. According to these theories, the risks that a programmer is obliged to prevent and that relate to their functional obligations (function-related risks) could be considered in principle, causally imputable. [117]

6. Conclusion

[112] S.C. Calvert, D.D. Heikoop, G. Mecacci, and B. van Arem, ‘A Human Centric Framework for the Analysis of Automated Driving Systems based on Meaningful Human Control’, 21 *Theoretical Issues in Ergonomics Science* 478 (2020), 490-492.

[113] Santoni de Sio and Van den Hoven, *supra* note 112, 9; *ibid*., 490 and 491 (emphasis added).

[114] *See* B. Boutin, ‘State Responsibility in Relation to Military Applications of Artificial Intelligence’ (forthcoming, 2022) (emphasis added).

[115] T. McFarland and T. McCormack, ‘Mind the Gap: Can Developers of Autonomous Weapons Systems be Liable for War Crimes?’, 90 *Int’l L. Stud*. 361 (2014), 366 (emphasis added).

[116] The anticipation of data issues is central to a recent UNIDIR report relating to data failures in AWs, Holland, *supra* note 45, 13 and 14.

[117] *See* Boutin and Woodcock arguing for the need to ensure MHC in the pre-deployment phase, *supra* note 10.

AV and AW are complex systems. Their programming implies a distribution of responsibilities and obligations within tech companies and among them and manufactures, third parties and users, which makes it difficult to identify who may be responsible for harm stemming from their use. Despite the temporal and spatial gap between the programming phase and crimes, the role of programmers in the commission of crimes should not be discarded. Indeed, crucial decisions on the *behaviour* and *effects* of AVs and AW are taken in the programming phase. While a more detailed case-by-case analysis is needed, this chapter has mapped out how programmers of AVs and AWs might be *in control* of certain AV- and AW-related risks and therefore criminally responsible for AV- and AW-related crimes.

By examining AV- and AW-related crimes, this Chapter has shown that the assessment of causation as a threshold for establishing whether an *actus reus* was committed may converge on the criteria of understandability and foreseeability of risks of unlawful *behaviour* and/or *effects* of AVs and AWs. It has been argued that those risks which fall within programmers' functional obligations and sphere of influence can be considered under their *control* and could be imputed to them.

Following this analysis, a notion of MHC applicable to programmers of AVs and AWs which is based on requirements for the imputation of criminal responsibility can be developed. As such, it may function as a responsibility 'anchor' or 'attractor' insofar as it helps trace back responsibility to the individuals that could understand, foresee, and anticipate the risk of a crime being committed with an AV or AW.

A final word of caution is warranted against the overcriminalisation of programmers' acts and omissions. Criminal responsibility is a last resort measure and is triggered by serious harm and a culpable mental state. Criminal law responses must always be weighed against the development of technology and its benefits – in this situation, the reduction of incident on the road and battlefield. In this context, consideration must also be given to whether civil liability, including product liability – which in Germany and the US has also been imported into criminal law[118] – and/or state responsibility for violations of IHL is instead best suited to address certain harm stemming from AVs and AWs. Finally, given the corporate and distributed nature of AI development and programming, corporate criminal responsibility – which is not accepted in all legal systems[119] – could also prove to be a viable avenue for accountability.

.

[118] *See* S. Gless *et al*, *supra* note 100, 426-429.

[119] For a comparative analysis, *see* K.J. Heller and M.D. Dubber, *supra* note 13.